University of Alberta

A Bayesian Analysis of Complex DNA Substitution Models

by

Ligia M. Mateiu

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Medical Sciences - Medical Genetics

Edmonton, Alberta
Spring 2007

Library and Archives Canada

Published Heritage Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et Archives Canada

Direction du Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 978-0-494-29711-7
Our file Notre référence
ISBN: 978-0-494-29711-7

NOTICE:
The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

AVIS:
L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l'Internet, prêter, distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

L'auteur conserve la propriété du droit d'auteur et des droits moraux qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.

Canada

Abstract

I have developed a method for calculating sequence substitution probabilities using Markov chain Monte Carlo (MCMC) methods. As a basic strategy, I used Uniformization to transform the original continuous time Markov process into a Poisson substitution process and a discrete Markov chain of state transitions. An efficient MCMC algorithm for evaluating substitution probabilities by this approach using a continuous gamma distribution to model site-specific rates is outlined. The method is applied to the problem of inferring branch lengths and site-specific rates from nucleotide sequences under a general time reversible (GTR) model and a computer program BYPASSR is developed. The method is applied to several viral datasets (HIV 1 *pol*, Japanese Encephalitis Virus genome and Lyssavirus glycoprotein), class I Major Histocompatibility Complex and RNases EDN/ECP. A large dataset consisting of 688 sequences of cytochrome *b* in mammals is also analyzed. In general, in protein coding sequences, the pattern expected for regions under negative selection is observed with third codon positions having the highest inferred rates, followed by first codon positions and with second codon positions having the lowest inferred rates. Several sites show exceptionally high substitution rates at second codon positions, which may represent the effects of positive selection.

I have also developed a novel method to address the errors occurring in the process of amplification of ancient DNA template, for which there is very limited methodology. The method is evaluated using simulated data sets under the new model. The correct identification of the nucleotide substitutions attributed to amplification errors validates the method.

Table of Contents

1 Introduction 1
- 1.1 Evolutionary rates . . . 1
- 1.2 Substitution models . . . 2
  - 1.2.1 DNA substitution as a Markov process . . . 3
  - 1.2.2 Codon and amino acid substitution models . . . 4
- 1.3 Models of substitution rate variation . . . 6
  - 1.3.1 Rate variation across sites . . . 6
  - 1.3.2 Rate variation across branches . . . 8
  - 1.3.3 Rate variation among sites and branches . . . 9
- 1.4 Calculating transition probabilities . . . 9
- 1.5 Likelihood calculation on a phylogenetic tree . . . 12
  - 1.5.1 Tree likelihood calculation via the pruning algorithm . . . 13
  - 1.5.2 Tree likelihood calculation using data augmentation . . . 13
  - 1.5.3 Ancestral state reconstruction . . . 14
- 1.6 Uncertainty in ancient DNA data due to miscoding lesions . . . 16

2 Estimation of site-specific rates 18
- 2.1 Model description . . . 19
  - 2.1.1 Data and parameters in the model . . . 19
  - 2.1.2 Uniformization of the Markov substitution process . . . 19
  - 2.1.3 Modeling Rate Variation Among Sites . . . 20
  - 2.1.4 Tree likelihood calculation using uniformization and data augmentation in the Bayesian approach . . . 22
- 2.2 Implementation . . . 23
  - 2.2.1 Metropolis Hastings algorithm . . . 24
- 2.3 Statistical performance . . . 30
  - 2.3.1 Simulating data to evaluate the method . . . 30
  - 2.3.2 Comparison with maximum likelihood implementation in PAML 30
  - 2.3.3 The influence of topology on substitution rate estimates . . . 34
  - 2.3.4 Testing extreme distributions of substitution rates . . . 40
  - 2.3.5 Analyzing independence among sites . . . 46
  - 2.3.6 Influence of branch length prior on parameter estimates . . . 49
  - 2.3.7 Implementation of the pruning algorithm . . . 49

2.3.8 Assessing convergence 53

3 Site-specific rate variation in several genes **57**
3.1 HIV-I *pol* gene 58
3.2 MHC class I 65
3.3 Japanese encephalitis virus and 3-prime noncoding 70
3.4 Lyssaviruses Rabies virus Genotype I glycoprotein with 3-prime noncoding region 74
3.5 Mammalian cytochrome b 81
3.6 Reconstructing the ancestral sequence of EDN and ECP genes 87

4 Modeling uncertainty in the fossil data **96**
4.1 Model description 96
4.2 Implementation 97
4.3 Statistical performance 99

5 Conclusions **107**

Bibliography **109**

A BYPASSR v.1.0 **119**
A.1 Overview 119
A.2 Installation notes 119
A.3 Input Files 120
A.3.1 Data file 120
A.3.2 Tree file 121
A.4 Control file **controlfile** 121
A.4.1 BYPASSR 121
A.4.2 Control file **BYPASSR-Gen** 126
A.5 Quick Start 128
A.6 Output Files 128
A.6.1 Burn-in 128
A.6.2 Sampling 129

List of Figures

1.1 Gamma distribution with different values of α. The mean of the distribution is α/β and the variance α/β^2. 7

1.2 The approximate "discrete" gamma distribution. The continuous distribution is split at the cut-off points for obtaining, in this case, four categories with equal probabilities. The vector of rates $\{r_1, r_2, r_3, r_4\}$ are the means (or median) rates for the four categories and represent the values of the rates the sites in each category can assume. 8

1.3 A phylogenetic tree, representing the data at a single site 14

2.1 Exponential (A) and uniform (B) densities of branch lengths prior. . . . 27

2.2 Comparison of estimated site-specific rates (panel A) and branch lengths (panel B) obtained using BYPASSR and BASEMLG programs for 6 taxa and 1000 sites dataset generated and analyzed assuming KHY85+Γ model. 31

2.3 Comparison of estimated site-specific rates (panel A) and branch lengths (panel B) obtained using BYPASSR and BASEMLG programs for 7 taxa and 1000 sites dataset generated and analyzed assuming JC+Γ model. 32

2.4 Plot of mean posterior site-specific rates from the BYPASSR program versus true rates. Data for the plot were simulated using either 10 taxa (panels A and B) or 50 taxa (panels C and D) and either 500 sites (panels A and C) or 5000 sites (panels B and D) under a GTR model. 33

2.5 Plot of empirical Bayes estimates of rates obtained using the BASEML program (with 20 rate categories for the discrete gamma approximation) versus true rates. Data for the plot were simulated using either 10 taxa (panels A and B) or 50 taxa (panels C and D) and either 500 sites (panels A and C) or 5000 sites (panels B and D) under a GTR model. 34

2.6 Plot of posterior distribution of site-specific rates for simulated data analyzed using the BYPASSR program. Panels A and B show posterior distributions for different numbers of taxa with an actual rate of 0.35 (indicated by vertical line) and either 500 sites (panel A) or 5000 sites (panel B). Panels C and D show the posterior distributions when the actual rate is either much higher, $r = 2.17$ (panel C) or much lower, $r = 0.09$ (panel D). 35

LIST OF FIGURES

2.7 Posterior mean of first 150 sites using the correct tree (tree 1, black line) and 9 random trees (label 2-10 in legend) for the dataset of 75 sequences and 1000 sites generated with $\alpha = 0.5$ and the GTR DNA substitution model. 38

2.8 Posterior distribution of substitution rate of a few selected sites at which the random topologies gave different estimates. Black vertical line corresponds to the distribution of substitution rate at that site obtained using the correct topology while distributions marked from 2 to 10 are for the substitution rate at site obtained using an incorrect topology. Dataset of 75 taxa and 1000 sites were generated with $\alpha = 0.5$ and analyzed using a GTR model. 39

2.9 Posterior distribution of rate at two random sites binned to calculate 95% highest posterior density. The bins with frequency less than 0.015 (panel A) and 0.004 (panel B) are not included in the HPD credible set. 41

2.10 95% Confidence intervals for posterior mean of substitution rates for the first 100 sites in the dataset of 25 (panel A) and 100 sequences (panel B). 43

2.11 Correlation coefficient and regression equations calculated for the true rates and the posterior mean of rates of BYPASSR (A panels) and marginal rates of BASEML (B panels). Datasets of 25, 50, 75 and 100 sequences (from top to bottom) and 1000 sites having substitution rates chosen from a continuous gamma distribution with $\alpha = 0.1$. 44

2.12 Boundaries for the 50 categories used by BASEML to approximate the gamma distribution with $\alpha = 0.1$. 45

2.13 BYPASSR and BASEML (50 cat.) branch lengths estimates for 25 sequences. True branch lengths are shown in black. 45

2.14 Correlation coefficient and regression equations calculated for the true rates and the posterior mean of rates of BYPASSR (A panels) and marginal rates of BASEML (B panels). Datasets of 25, 50, 75 and 100 sequences (from top to bottom) and 1000 sites having substitution rates chosen from a continuous gamma distribution with $\alpha = 2.5$. 47

2.15 Posterior distribution of the rate for a typical site when $\alpha = 2.5$ when the number of sequences vary from 25 to 100. The black line shows the prior distribution, Gamma with parameter $\alpha = 2.5$. 48

2.16 Correlation matrix for the 37 sites with mean substitution rate > 4 (panel A) and for the 134 sites with mean rate < 0.1 (panel B). 49

2.17 Trace plots of the α, the tree length and the marginal tree log likelihood, for the dataset of 20 taxa and 500 sites, as a function of run time with equal run time using either a pruning algorithm or data augmentation with the uniformization technique. 51

2.18 Trace plots of α and tree length for two datasets with either 20 or 50 sequences and 2000 sites. Equal run time is used for both datasets analyzed under JC69 substitution model with either pruning algorithm or data augmentation with the uniformization technique. 52

2.19 Calculating the degree of overlapping of substitution rate distribution at two sites. The area inside the gray zone gives the value of R. 54
2.20 R during burn-in for a dataset of 45 sequences and 750 sites. 55
2.21 Tree length during burn-in for a dataset of 45 sequences and 750 sites. . 55
2.22 Tree log Likelihood during burn-in for a dataset of 45 sequences and 750 sites. 56

3.1 Sites (and their codons) with $\overline{r_1} < \overline{r_2} > \overline{r_3}$ and posterior probability P($r_2 > 1$)> 0.5 (panel A). Sites with $\overline{r_1} > r_3$ and > 0.9 P($r_1 > 1$)> 0.9 (panel B). The codons in italic show the posterior probability > 0.95 of either $r_2 > 1$ (top panel) or $r_1 > 1$. 61
3.2 Posterior distribution of the substitution rate at site 8, 134, 938 and 2312 when the maximum likelihood tree (line 1 in the legend) and 8 random trees (lines from 2 to 9) are used to analyze HIV data set. 64
3.3 Comparison of estimated branch lengths and site-specific rates obtained using BYPASSR and BASEML programs. Panels plot the mean site-specific rates from the posterior distribution generated by BYPASSR (horizontal axis) against estimates of site-specific rates generated using BASEML (vertical axis) with either 5 (panel A), 20 (panel B) and 50 (panel C) rate categories, respectively. 66
3.4 Codons in MHC class I gene with $\overline{r_1} < \overline{r_2} > \overline{r_3}$ and posterior probability P($r_2 > 1$) > 0.8 (panel A). Sites with $\overline{r_1} > \overline{r_3}$ and P($r_1 > 1$) > 0.8 (panel B). The substitution rates are represented in expected numbers of substitutions per unit time. 68
3.5 Codons in MHC class I gene having posterior probability P($\overline{r_2} > 1$) ≥ 0.8 (panel A). Sites with posterior probability P($r_1 > 1$) ≥ 0.8 (panel B). The substitution rates are represented in expected numbers of synonymous substitutions per unit time. 69
3.6 Phylogenetic tree of 20 isolates of Japanese Encephalitis complete genome. 72
3.7 Distribution of substitution rates estimates in BYPASSR and BASEML with 5, 20 and 50 categories for Japanese encephalitis virus genome. . . 73
3.8 Maximum likelihood tree for 35 isolates of lyssavirus. 75
3.9 Comparison of estimated branch lengths and site-specific rates obtained using BYPASSR and BASEML programs. Left Panels A, C, E plot mean branch lengths from the posterior distribution generated by BYPASSR (horizontal axis) against estimates of branch lengths generated using BASEML with 5, 20 and 50 rate categories (vertical axis). Right Panels B, D and F plot the mean site-specific rates from the posterior distribution generated by BYPASSR (horizontal axis) against estimates of site-specific rates generated using BASEML (vertical axis) with either 5, 20 or 50 rate categories, respectively. 77

3.10 Mean posterior rate at codon positions of glycoprotein gene and at the sites located in the 3-prime noncoding region. The approximate location along the gene of the signal peptide (SP), endodomain, ectodomain, transmembrane domain (TM), antigenic sites and Western Blot positive epitope (open arrow heads) is shown as in [1]. 79
3.11 Mean posterior distributions at first, second and third codon positions of signal peptide and transmembrane domains (A) and endodomain (B). The sites located at the second position or first codon position that are good candidates to evolve under positive selection are marked with ✦ and ✧, respectively. Panel C shows the mean posterior rates for sites in the 3-prime noncoding region. 80
3.12 Chain convergence during the 9 burn-in runs and sampling for α Panel (A), log L tree (B) and tree length (C) for cytochrome *b* dataset 83
3.13 Correlation matrix for the first 100 sites (left panel) and for the last 100 sites (right panel). 84
3.14 Posterior distribution at some sites with high rates. 84
3.15 Mean posterior substitution rates at candidate sites for positive selection in cytochrome *b* gene. Panel A shows the codons that have $\overline{r_3} < \overline{r_2} > \overline{r_1}$ and $\overline{r_2} > 3$ with probability >0.95 in italics. The codons with $\overline{r_1} > \overline{r_3}$ and $r_1 > 3$ with probability >0.95 (italics) are in the bottom panel. . . . 85
3.16 Mean posterior substitution rates at candidate sites for positive selection in cytochrome *b* gene. Top panel show the codons that have $\overline{r_3} < \overline{r_2} > \overline{r_1}$ and $2 < \overline{r_2} < 3$ with probability >0.95 in italics. The codons with $\overline{r_1} > \overline{r_3}$ and $2 < \overline{r_1} < 3$ with probability >0.95 (italics) are in the bottom panel. 86
3.17 Phylogenetic tree of the EDN/ECP gene for 18 primate sequences. [2]. . 92
3.18 Protein reconstruction at internal nodes A and B as inferred with parsimony, Bayesian and maximum likelihood methods 93
3.19 Mean posterior substitution rates at first, second, and third codon positions of the EDN/ECP gene. The possible nucleotide sites found by BYPASSR to evolve under positive selection (❶). The codons found by CODEML to be positively selected and also met criteria of containing a positively selected nucleotide site in BYPASSR (❷); The extra codons found by CODEML (❸). 95

4.1 Phylogenetic tree under degradation model. 97
4.2 Algorithm for root node (left) or internal node sliding (right) for molecular clock implementation. 99
4.3 Generate data for degradation model testing. 100

4.4 Comparison between the mean posterior substitution rates when degraded data (+) is analyzed with BYPASSR-degr and BYPASSR (A panels). B panels show the correlation between the mean posterior rates obtained from degraded data using BYPASSR-degr versus original data set before including the errors (-) analyzed with BYPASSR. The simulated data sets have 20, 30, 40 and 50 sequences (from top to bottom). (set 4 in Table 4.1). 105

4.5 Comparison between the mean posterior substitution rates when "degraded" data (+) is analyzed with BYPASSR-degr and BYPASSR (A panels). B panels show the correlation between the mean posterior rates obtained from degraded data using BYPASSR-degr versus original data set before including the errors (-) analyzed with BYPASSR. The simulated data sets have 50, 60, 70 and 80 sequences (from top to bottom). The proportion fossil/extant sequences is equal (Table 4.3). 106

Chapter 1

Introduction

DNA mutations are the ultimate source of variation upon which evolution acts. Interspecies and intra species genetic variation caused by nucleotide substitutions, insertions, deletions, recombination, gene conversion, etc. can be a reflection of the effects of natural selection, random genetic drift, or both. Mutations that do not impair the reproduction of an organism may become fixed in a population and, if no other mutation occurs, are transmitted to descendants. The history of a sample of extant (or fossil) sequences can be inferred by building a phylogenetic tree based on a model that describes the mechanism of DNA substitutions. As the mutations that affect the genome evolution on a long-term basis are those that become fixed as nucleotide substitutions, my research is focused on understanding and modeling the nucleotide substitution process and identifying potentially relevant sites (e.g., those sites under selection) from an evolutionary perspective.

1.1 Evolutionary rates

Rates of DNA sequence change are determined by the nucleotide mutation rate and the subsequent effects of natural selection. Normally, most mutations occur as copying errors during DNA replication that escape the repair mechanisms. The base composition of a sequence also influences the mutation rate, as shown by the increased frequency of transitions versus tranversions over that expected if all substitutions were equally likely. Another factor influencing the frequency of substitutions is their location in the genome. Substitutions that occur in coding regions of DNA are typically reduced in number by comparison with non-coding regions, probably due to the effects of negative selection against deleterious mutations. Particular importance is attributed to mutations in the coding DNA. A mutation that does not change the amino acid sequence of a protein is called a silent or synonymous mutation. If a change in the sequence of the gene product occurs, the mutation is said to be nonsynonymous. However, some mutations in the noncoding DNA are also of functional importance. Mutations in regulatory regions or important intronic locations might change the gene expression of the regulated gene.

Abundant evidence also suggests substitution rates differ between loci. Many other factors such as body size, metabolic rates, generation time [3], and different CpG content [4], also influence the evolutionary rate of species or genes. However, when individuals of the same species or closely related species are analyzed, the substitution rate of a given gene is often found to be similar. Constant evolutionary rates over time have been referred to as the molecular clock [5].

Although it is not yet clear how to differentiate between variable mutation rate and selection at the nucleotide level, some indications exist. For instance, the substitution rate varies according to the positions occupied by the sites within a codon: the site at the third codon position has the highest rate followed by the first codon position and then the second position. The nature of the genetic code specifies that only a few third-position substitutions lead to amino acid substitution. The substitution rates at such sites are considered more reflective of mutation than selection. On the other hand only a small proportion of second-position and first-position substitutions are silent and the contribution of selection in determining the substitution rate at such sites is likely more significant.

A very conserved gene has a high percent of sequence identity among distant related species. These genes are characterized by having an extremely low rate of nonsynonymous codon substitution compared with other genes. However, even a gene that is highly conserved may have a few sites that are not so critically necessary in maintaining its structure and function. Evolutionary forces might occasionally modify the nucleotide at such sites and potentially improve the function of a conserved gene.

Theoretical analyses of the variability of evolutionary rates have revealed important aspects of evolution by successfully identifying genes or sites under selection. These new statistical methods provide important results that have helped the experimentalists to focus on evolutionary relevant parts of a gene or genome. On the other hand, biologists provide a continuous stream of questions which new theoretical tools are needed to answer. This sort of cooperation has resulted in strong evidence that variable selection pressures exist at the molecular level. For example, evidence of positive selection has been found by investigating rate variation among sites in many categories of genes, predominantly genes involved in immunological responses, reproductive function and digestion.

1.2 Substitution models

Although many types of mutations occur in DNA and become fixed in genomes, base substitutions are the type most often analyzed in phylogenetic studies. The DNA substitution process is very complex, only partially understood, and can only be modeled in probabilistic terms. The probabilistic model, a collection of probability density functions of variables, is a great simplification of reality, but it is the only currently available tool to investigate the principal components of DNA substitution and their interactions. A typical phylogenetic analysis has the topology, the branch lengths and the parameters

of the substitution model as the most relevant variables. Even with many assumptions that reduce the number of free parameters in the models an enormous number of parameters remain. In general, the more parameters, the better the model is in explaining the biological process. Particularly in the case of nested models, models that are more complex are generally favored [6].

The important influence that a misspecified DNA substitution model can have on the accuracy of many phylogenetic inference methods is now well established. The effects of overspecifying versus underspecifying a model can be very different, however. For example, recent simulation studies suggest that Bayesian posterior probabilities of phylogenetic trees can be inflated if an overly simple substitution model is used, while an overly complex model can produce accurate posterior probabilities if the true model is a submodel [6] [7]. Many studies have shown that taking account of among-site rate variation, in particular, is very important for obtaining accurate point estimates of phylogeny and branch lengths [8], as well as accurate posterior probabilities for trees [6]. The influence of tree topology on estimates of rate variation among sites is not fully clarified. A relatively accurate topology was found to suffice in most of the studies [9] [10]. On the other hand, a wrong or underspecified substitution model that ignores among-site rate variation significantly affects the topology and branch length estimates. From a biological perspective, parameter-rich substitution models may lead to new patterns and insights that would be missed using a simpler model. For example, allowing the ratio of nonsynonymous (dN) to synonymous (dS) substitution rates to vary across codons in a gene can highlight important residues that have been under positive or negative selection [11] and such phenomena would be missed using a model that ignores among-site variation in the substitution process. It is becoming evident that more realistic parameter-rich substitution models should be used for phylogenetic inference, especially in the Bayesian framework. Commonly used models are known to be too simple and often fit sequence data poorly [12] [13].

1.2.1 DNA substitution as a Markov process

It has become standard for any phylogenetic analysis involving a limited number of sequences to use an explicit model to describe the DNA substitution process. A stochastic process, called a Markov process, is commonly used. A Markov process is described in terms of conditional probabilities: the conditional probability distribution of future states of the process, given the present state, is conditionally independent of the past states. Considering a nucleotide as the evolutionary unit, the rates of change from one nucleotide to another during an infinitesimal amount of time are given by a 4x4 instantaneous matrix, Q [14].

$$\mathbf{Q} = \{q_{ij}\} = \begin{pmatrix} q_{11} & a\pi_{\mathrm{C}} & b\pi_{\mathrm{A}} & c\pi_{\mathrm{G}} \\ a\pi_{\mathrm{T}} & q_{22} & d\pi_{\mathrm{A}} & e\pi_{\mathrm{G}} \\ b\pi_{\mathrm{T}} & d\pi_{\mathrm{C}} & q_{33} & f\pi_{\mathrm{G}} \\ c\pi_{\mathrm{T}} & e\pi_{\mathrm{C}} & f\pi_{\mathrm{A}} & q_{44} \end{pmatrix}, \tag{1.1}$$

where a, b, c, d, e, f are the relative rate parameters and π_T, π_C, π_A and π_G are the nucleotide frequency parameters at stationarity (e.g. the long-run frequencies of the bases T, C, A and G) [15]. The elements on diagonal q_{ii} are the negative sum of the terms on the corresponding row (e.g. $q_{11} = -(a\pi_C + b\pi_A + c\pi_G)$, $q_{22} = -(a\pi_T + d\pi_A + e\pi_G)$, etc). This matrix corresponds to the most general time-reversible model, the so called general time reversible model (GTR) [16]. Time reversibility implies the same rate of change from the nucleotide i to j as from j to i at stationarity. The mathematical advantage of imposing the reversibility condition consists in allowing the likelihood calculation on a phylogenetic tree to be independent of the root location [14].

All other substitution models are special cases of GTR. The simplest DNA substitution model, JC69, dates back to 1969 and is the work of Thomas Jukes and Charles Cantor [17]. Under this model, the base frequencies and the relative rates of substitution are equal. Intermediate models allow nucleotide frequencies or some of the relative rates to vary. For example, Kimura [18] proposed a model in 1980 that constrains stationarity, base frequencies to be equal, but allows different rates for transitions and transversions (see Swofford *et al.* for review [19]).

1.2.2 Codon and amino acid substitution models

Substitution models may use the amino acid as the evolutionary unit instead of the nucleotide if the sequences are protein coding. The transition probabilities from a particular amino acid to any other amino acid are also then described as a Markov process, but the matrix Q is a 20x20 matrix. If the time-reversibility condition is applied, the equivalent of a GTR matrix for amino acids would have 190 off-diagonal parameters. To avoid such a large number of free parameters, the vast majority of the amino acid models are empirical. The entries in the instantaneous transition rate matrix are obtained by averaging over some collections of amino acid data sets. The most common empirical matrices are PAM [20], BLOSUM [21], JTT [22], WAG [23]. An obvious preference for using amino acid data instead of nucleotides is given by a dataset with a high variation in the base frequencies among species, which causes a reduction in the phylogenetic information [24] [25].

Codon based substitution models, originally formulated in a maximum likelihood framework, [26], [27] consist of a 61x61 matrix, with stop codons excluded. A typical model based on codon triplets is parametrized such that the number of free parameters is reduced. The substitution rates from codon i to codon j in the basic codon based model [28] take the values

$$q_{(i,j)} = \begin{cases} 0 \text{ if the two codon differ at more than one position} \\ \pi_j \text{ for synonymous transversion} \\ k\pi_j \text{ for synonymous transition} \\ \omega\pi_j \text{ for nonsynonymous transversion} \\ \omega k\pi_j \text{ for nonsynonymous transition} \end{cases}$$

The only free parameters are a transition/transversion ratio k and the nonsynony-

mous/synonymous rate ratio ω. Usually, the nucleotide frequencies at each codon position are considered to be at stationarity and estimated from the data [29] [30]. The codon based models are preferred when the objective is to detect selection because of the direct link between ω and selection pressure. The interpretation of the ω value is as follows: a site with $\omega > 1$ means a greater chance for the nonsynonymous changes to become fixed in the population relative to the synonymous changes at the site and this is, in general, considered to be the signature of positive selection; $\omega < 1$ indicates a slower accumulation of nonsynonymous substitutions than synonymous ones, presumably because of their deleterious effect; $\omega = 1$ indicates neutrality, giving equal chance that both types of changes become fixed. However, there are situations when the dN/dS ratio is not capable of identifying positive selection. The parameter ω is an average measure of selection over time and over sites. In a gene with large proportion of conserved sites, but a few fast evolving sites the overall ω will not be significantly greater than 1 [29]. All the models aimed at identifying selection through ω assume a constant synonymous substitution rate across sites with ω representing only the variation in nonsynonymous sites [31]. The same ω is also assigned to all nonsynonymous changes, which is not necessarily true. The proposed distribution of ω across sites is not limited to the gamma or log-normal distribution, as are the substitution rates in nucleotide models.

Another type of codon model partitions the data into codon positions. Such models use nucleotide substitution models, but the information contained in the genetic code is also considered. The relevant parameters such as tree topology and substitution matrix with the relative rates and nucleotide frequencies can be shared or specific to the codon partitions, generating various combinations with different number of free parameters. If rate variation among codon positions is assumed, three gamma distributions with different parameter α are defined. The same parameterization is extended by Yang (1996) [32] to datasets containing multiple genes, treating a gene as a codon position. Shapiro *et al.*(2006) [33] compared a large range of nucleotide substitution models with "classic" codon models and the mixture of the two. The model based on data partitioning was found to be more computationally efficient in comparison with the codon models eliminating the large codon substitution matrix. Moreover, the improved biological reality makes them suitable for protein coding sequences.

There are interesting applications of substitution models to gene coding data (e.g., the use of ω as a measure of the strength of selection, the use of amino-acid biological and physiochemical properties etc.), but coding DNA is overall a very small part of the genome. The remaining non-coding regions in the genome are also crucial in understanding molecular evolution [34] [35] and DNA substitution models rather than codon models must be used. The sequence conservation in some non-coding regions was found to be remarkable [36], but sometimes the alignment of noncoding DNA sequences is problematic. However, the alignment difficulty can be overcome by the availability of sophisticated programs that perform such tasks [37] [38].

Wong and Nielsen [39] proposed a statistical method to search for selection in noncoding regions, based on the assumption of a constant rate of evolution in both the coding and noncoding regions of the gene. Their method is intended for use with

multiple sequence alignments containing both coding and noncoding sequences. The coding segment is analyzed according to the method of Yang *et al.*(2000) [29] that identifies positive selection at codon sites with ω greater than one. The noncoding region of the alignment is analyzed following the Hasegawa-Kishino-Yano DNA substitution model [40] with an additional parameter that connects the coding and noncoding regions. More precisely, each term of the instantaneous substitution matrix is multiplied by a substitution rate normalized by the synonymous nucleotide substitution rate in the coding region, maintaining the same biological interpretation as for ω. In their analysis of 13 viral datasets, little or no evidence of positive selection was found in the noncoding regions.

1.3 Models of substitution rate variation

1.3.1 Rate variation across sites

The assumption (implicit in the work of Jukes and Cantor and others) that nucleotide substitutions at each site of a DNA sequence follow a Poisson process with a common rate conflicts with biological observations suggesting rates vary across sites [41] [42]. Historically, the gamma [43] [44] [45] [46] and lognormal [47] prior distributions were used to model among-site rate variation to compute genetic distances. Yang (1993) [15] corrected the maximum likelihood method for phylogenetic inference to allow for rate variation across sites by modeling rate variation using a Gamma distribution [15] [48]. Frequent use of the Gamma distribution is justified by mathematical tractability (positive values ranging from 0 to infinity, flexible shape of the distribution, parameters for mean and variance) rather than biological motivation. Following Yang (1993), the density for rates r for m sites is a continuous gamma:

$$f(r_m|\alpha) = \frac{\alpha e^{-\alpha r_m}(\alpha r_m)^{\alpha-1}}{\Gamma(\alpha)}.$$

The value of α is inversely proportional to the degree of the rate variation. A value of $\alpha \leq 1$ indicates a high rate variation among sites with the vast majority of rates close to 0, while a few rates have extremely high values. The gamma distribution takes the shape of a symmetrical modal distribution as α increases meaning that most of the sites have intermediate rates (see Fig. 1.1).

Time and rate are confounded parameters in studying levels of divergence among sequences and only their product can be estimated. A mathematically convenient further parameterization of the Gamma distribution is to set its mean at one, by allowing the shape parameter α to be equal to the inverse of the scale parameter β [15]. This parameterization sets the average relative substitution rate over all sites of each dataset to be 1.

Formally written, the likelihood of a site is

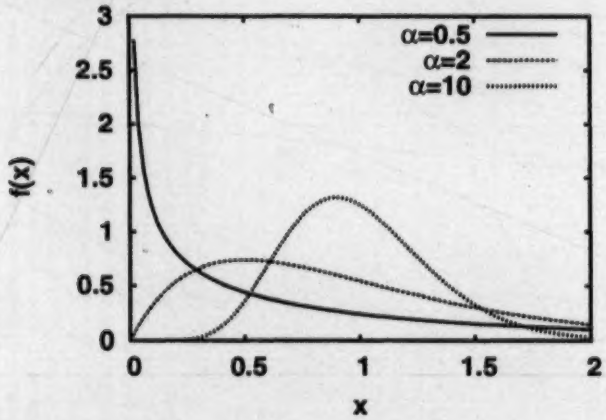


Figure 1.1: Gamma distribution with different values of α. The mean of the distribution is α/β and the variance α/β^2.

$$L = \int_0^\infty f(r)L(\tau, r)dr, \tag{1.2}$$

where $L(\tau, r)$ is the likelihood of a site for tree τ with rate r and $f(r)$ is the density function of the distribution of rates [49]. The calculation of $L(\tau, r)$ involves a very large number of terms even for only a few sequences. When the dataset has more than a few species, evaluation of the likelihood is hampered by computational difficulties. To address this problem, Yang (1994) [50] proposed an approximate method that discretized the gamma distribution [50]. The continuous distribution is split into several categories with equal probabilities and the mean of each category gives the rate for all the sites within the category (Fig. 1.2). This approximation is commonly used in the available software for phylogenetic inference. By applying this approximation, equation 1.2 becomes

$$L = \sum_{i=1}^{n} p_i L(\tau, r_i) \tag{1.3}$$

The p_is are the probabilities (weights) of the ith quantile. Including the site specific rate in equation Eq. (1.5) this becomes,

$$P(t) = e^{Qtr_i}. \tag{1.4}$$

The matrix of probabilities of a change to any nucleotide from another nucleotide over time period t at the ith site of the DNA sequence alignment is given by exponentiating the product Qtr_i, where r_i is the relative substitution rate at the ith site. In phylogenetic analysis, the Q and t are confounded parameters (cannot be estimated independently of

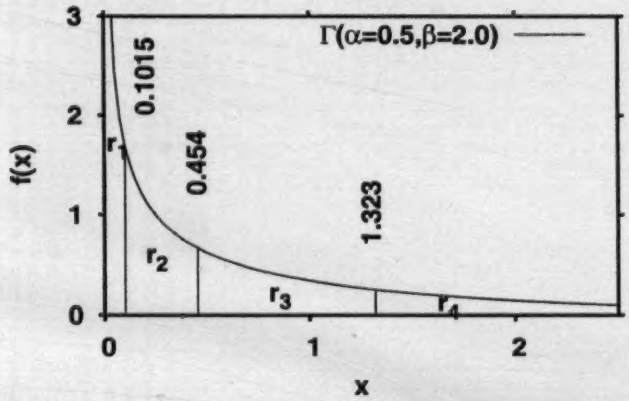


Figure 1.2: The approximate "discrete" gamma distribution. The continuous distribution is split at the cut-off points for obtaining, in this case, four categories with equal probabilities. The vector of rates $\{r_1, r_2, r_3, r_4\}$ are the means (or median) rates for the four categories and represent the values of the rates the sites in each category can assume.

one another) and usually the matrix Q is scaled such that the mean number of substitution per unit time is equal to 1 [15]. The scaling is done by multiplying the matrix Q by a normalizing constant.

1.3.2 Rate variation across branches

In 1962, Emile Zuckerkandl and Linus Pauling compared the amino acid sequences of hemoglobins from different species and noticed that genes appeared to evolve at a relatively constant rate. From this observation, they generalized and formulated the molecular clock theory. The graphical representation for a phylogenetic tree under the molecular clock is given by a rooted tree with branch lengths proportional to expected numbers of substitutions and in which all terminal taxa are equidistant from the root. The phylogenetic implication of this theory is that the amount of time since species diverged can be estimated simply by comparing their gene sequences. However, the molecular clock assumption typically holds only for sequences that come from very closely related species [51] [52]. As well, the rates of substitution (and thus the calibration of the clock) varies greatly across different genes.

Phylogenetic methods dealing with rate variation across lineages allow various degrees of relaxation of the molecular clock assumption. A global molecular clock implies a constant substitution rate on the whole tree. A local molecular clock implies that a part of the tree is evolving at a constant rate, while the rest of the branches have evolutionary rate variation. The most general model would allow each branch to have a freely varying

substitution rate. Because time and rate are confounded parameters in phylogeny and only their product can be estimated, one can imagine an equivalent process with a constant rate across branches and freely varying branch lengths. Expressing this in terms of the rate matrix Q, as long as Q is shared by all the branches, the expected branch lengths are allowed to depart from the molecular clock (e.g. [14]). This is the common approach in most of the programs and methods for phylogenetic inference, when a molecular clock is not imposed.

1.3.3 Rate variation among sites and branches

Evolution is expected to change the substitution rate of some sites along a few branches in a phylogenetic tree. In other words, some sites might evolve rapidly in some parts of the tree but slowly in other parts. The earliest model of this form is attributed to Fitch [53] who called it the "covarion" model. The most straightforward way to model rate variation among sites and along branches would be to allow each site to have its own substitution matrix and each branch to have its own transition probabilities. Because of the complexity of such a model, theoreticians have had to resort to more simplified versions. An example of such a model is the branch-site model of Zhang *et al* [54]. The authors have a set of prespecified branches that evolve with one "foreground" rate and the rest of the branches evolve at another "background" rate. The codon sites along the molecule are assumed to be in four classes: conserved over the entire tree, neutral evolution [55] throughout the tree, conserved or neutral on background branches, but positively selected on foreground branches.

1.4 Calculating transition probabilities

The rate matrix Q describes substitution rates during an infinitesimal time interval dt. If one wants to calculate transition probability along a branch of a phylogenetic tree of duration t, the matrix Q has to be exponentiated,

$$P(t) = e^{Qt}. \tag{1.5}$$

The P matrix is called the transition probability matrix.

Matrix exponentiation using matrix decomposition

For the JC69, K80, HKY85 and other similar models, P can be calculated by using the power series, as follows:

$$P(t) = e^{Qt} = I + Qt + \frac{1}{2!}(Qt)^2 + \frac{1}{3!}(Qt)^3..., \tag{1.6}$$

where I is the identity matrix with value 1 for diagonal and 0 for off-diagonal elements. Each term in the sum can be written as a function of the previous terms and the whole

sum can be reduced to a simple analytic solution. The powers of the Q matrix of the GTR model cannot be easily analytically deduced, however, and a numerical approach is therefore required. First, matrix diagonalization is generally used to find the eigenvalues and the eigenroots of Q. Matrix Q is then decomposed into a product of three matrices.

$$Q = HDH^{-1}, \tag{1.7}$$

where H is the matrix that has on its columns the eigenvectors and H^{-1} is the inverse of H. The D matrix contains on its diagonal the eigenvalues $\{D_1, D_2, D_3, D_4\}$ and 0 for off-diagonal elements.

The transition matrix P is obtained by replacing the eigenvalues $\{D_1, D_2, D_3, D_4\}$ by their exponentials in the matrix product.

$$P(t) = H \begin{pmatrix} e^{D_1 t} & 0 & 0 & 0 \\ 0 & e^{D_2 t} & 0 & 0 \\ 0 & 0 & e^{D_3 t} & 0 \\ 0 & 0 & 0 & e^{D_4 t} \end{pmatrix} H^{-1} \tag{1.8}$$

Uniformization technique

An alternative to matrix exponentiation for obtaining the elements of $P(t)$ is the uniformization (randomization) technique [56] [57]. The idea is to rewrite the continuous time Markov chain as a probabilistically identical process in which discrete state changes occur at random times to facilitate the calculation of the transition probability matrix $\mathbf{P}$ in equation 1.5.

The continuous time Markov process is described by the instantaneous rate matrix $\mathbf{Q}$ and the diagonal terms q_{ii}, $i = \{1, 2, 3, 4\}$ are the exponentially distributed waiting times in i state. That is, once in state i, the process leaves the current state with the exponential rate q_{ii}. Furthermore, when a move occurs, it is to state j with probability q_{ij}/q_{ii}.

An equivalent process can be found by allowing transitions to occur at discrete points. The new discrete process has the same state space as the continuous time process, but the waiting time between events is exponentially distributed with a rate ν independent of the current state of the process. That is, all states have exponentially distributed waiting times with the same parameter ν. The probability of moving from a state i to state j is q_{ij}/ν. The rate ν has to be greater than any $q_{\{ii\}}$ to ensure that there is at least one single-step transition from a state to itself (the chain is aperiodic). The number of transitions in the discrete time Markov chain is then Poisson with rate ν and the desired equivalent process is obtained.

More formally, the one step transition probability matrix under the uniformized process, $\mathbf{P}$, is

$$\mathbf{P} = I + \frac{\mathbf{Q}}{\nu}. \tag{1.9}$$

where I is the identity matrix, ν is the rate of the process and $\nu > max_i q_{ii}$.

The normalized (on rows) matrix $\mathbf{P}$, is the discrete Markov chain associated with a Poisson process conditional on the number of transitions at time t,

$$P(t) = \sum_{M=0}^{\infty} \frac{(\nu t)^M e^{-\nu t}}{M!} \times (P)^M, \tag{1.10}$$

where M is the number of substitutions under the uniformized process and $(P)^M$ is the discretized process under uniformization raised to the M*th* power. In other words, the transition probabilities are obtained by summing over the product of the discrete transition probability given M events and the probability of M events under the uniformized process, the M-step transition probability matrix.

Example: Uniformization of a two-state Markov process

To illustrate the uniformization procedure in a concrete case, I consider a simple two-state continuous time Markov process with instantaneous rate matrix,

$$\mathbf{Q} = \begin{pmatrix} -a & a \\ b & -b \end{pmatrix}.$$

Letting $\nu = a + b$, the Markov chain specifying the transition probability at each jump events is

$$\mathbf{P} = \begin{pmatrix} 1 - \frac{1}{\nu}a & \frac{1}{\nu}a \\ \frac{1}{\nu}b & 1 - \frac{1}{\nu}b \end{pmatrix} = \begin{pmatrix} \frac{b}{a+b} & \frac{a}{a+b} \\ \frac{b}{a+b} & \frac{a}{a+b} \end{pmatrix}.$$

In this particular example, all powers of the matrix are identical to the original (i.e., the matrix is idempotent) except, of course, the zero power which is the identity matrix. To calculate transition probabilities under this process, I marginalize by summing over the product of the discrete transition probability given M events and the probability of M events under the uniformized process,

$$p_{ij}(t) = \sum_{M=0}^{\infty} \frac{(\nu t)^M e^{-\nu t}}{M!} \times P_{ij}^M. \tag{1.11}$$

This simplifies to give,

$$\begin{aligned} p_{11}(t) &= \frac{b + ae^{-(a+b)t}}{a+b}. \\ p_{12}(t) &= \frac{a(1 - e^{-(a+b)t})}{a+b}. \\ p_{21}(t) &= \frac{b(1 - e^{-(a+b)t})}{a+b}. \\ p_{22}(t) &= \frac{a + be^{-(a+b)t}}{a+b}. \end{aligned} \tag{1.12}$$

Transition probabilities can be solved by exponentiating the matrix. The eigenvalues are 1 and $-(a+b)$, the matrix of right eigenvectors is

$$\mathbf{H} = \begin{pmatrix} 1 & -\frac{a}{b} \\ 1 & 1 \end{pmatrix},$$

and the inverse of $\mathbf{H}$ is

$$\mathbf{H}^{-1} = \begin{pmatrix} \frac{b}{a+b} & \frac{a}{a+b} \\ -\frac{b}{a+b} & \frac{b}{a+b} \end{pmatrix}.$$

If D is defined to be a matrix with diagonal elements that are the eigenvalues then

$$e^{Dt} = \begin{pmatrix} 1 & 0 \\ 0 & e^{-(a+b)t} \end{pmatrix},$$

and

$$\mathbf{H}e^{Dt}\mathbf{H}^{-1} = \begin{pmatrix} \frac{b+ae^{-(a+b)t}}{a+b} & \frac{a(1-e^{-(a+b)t})}{a+b} \\ \frac{b(1-e^{-(a+b)t})}{a+b} & \frac{a+be^{-(a+b)t}}{a+b} \end{pmatrix},$$

which agrees with the previous result obtained by uniformization of the process. In complex models, it is natural to use MCMC to evaluate the sum of Equation 1.11 and such a procedure will be used in developing an MCMC method that evaluates transition probabilities numerically.

In chapter 2, I will present a novel method for calculating transition probabilities using uniformization and data augmentation.

1.5 Likelihood calculation on a phylogenetic tree

The likelihood is the probability of observing the data given a particular model viewed as a function of the parameters given the data. Usually, the parameters to be estimated in phylogenetics are the tree topology τ, the branch lengths $\mathbf{v}$, the substitution matrix parameters θ and, in our case, also includes parameters of a model of site specific rate variation. Using various algorithms, a combination of parameters is found such that the likelihood function is maximized and this set of parameters is considered to best explain the data.

A common assumption in phylogenetic inference studies is that substitutions at different sites are independent. Although biological support for this model does not always exist, it is a very attractive hypothesis from the mathematical and computational points of view. In this way, the likelihood at each site is calculated along the tree and the total likelihood is a simple product of site likelihoods. The likelihood for the tree in Fig. 1.3 is the product of the likelihood on each branch times the probability of having nucleotide x at the root. The root location on tree τ does not influence the calculations as long as the substitution process is reversible, meaning that the direction of the process along the branches does not matter.

$$\begin{aligned}\text{Prob}(A,C,T|\tau) &= \sum_x \sum_y \text{Prob}(x)\text{Prob}(T|y,t_4)\text{Prob}(C|y,t_3) \\ &\quad \times \text{Prob}(A|x,t_1)\text{Prob}(y|x,t_2) \end{aligned} \tag{1.13}$$

1.5.1 Tree likelihood calculation via the pruning algorithm

In a tree with n species, equation 1.13 above has 4^{n-1} terms, which is a very large number if $n > 15$. The solution to this problem, developed by Joseph Felsenstein in 1981 [14], is the pruning algorithm. Considering a node k with left child l and right child m at the ends of the branches t_l and t_m, the conditional likelihood of subtree s is [58]

$$L_k^{(i)}(s) = \left(\sum_x \text{Prob}(x|s,t_l L_l^{(i)}(x))\right)\left(\sum_y \text{Prob}(y|s,t_m L_m^{(i)}(y))\right)$$

By rewriting equation 1.13 and moving the summation sign as far right as possible, it becomes

$$\begin{aligned}\text{Prob}(A,C,T|\tau) &= \sum_x \text{Prob}(x)\text{Prob}(A|x,t_1) \sum_y \text{Prob}(T|y,t_4) \\ &\quad \times \text{Prob}(C|y,t_3)\text{Prob}(y|x,t_2) \end{aligned} \tag{1.14}$$

The number of summation terms is greatly reduced and for large trees this leads to an important reduction in the computational time. However, every time a branch length is modified in the Eq. 1.14, the substitution matrix at that branch has to be recalculated (see Eq. 1.8) and all the sums involving the modified branch length have to be updated.

1.5.2 Tree likelihood calculation using data augmentation

A very different approach was taken by Robinson *et al.* [59] in a model for protein evolution in the context of phylogenetic inference. The complexity of the model is greatly increased because they did not use the common assumption of independent evolution of codons, which allows the instantaneous substitution matrix to be written as a 64×64 matrix. Instead, their substitution matrix, R, is of $4^N \times 4^N$ size, where N is the nucleotide sequence length, representing the rate of change from sequence i to sequence j:

$$\mathbf{R_{ij}} = \begin{cases} u\pi_h \text{ for a synonymous transversion} \\ u\pi_h k \text{ for a synonymous transition} \\ u\pi_h \omega e^{(E_s(i)-E_s(j))s+(E_p(i)-E_p(j))p} \text{ for a nonsynonymous transversion} \\ u\pi_h k\omega e^{(E_s(i)-E_s(j))s+(E_p(i)-E_p(j))p} \text{ for a nonsynonymous transition} \end{cases}$$

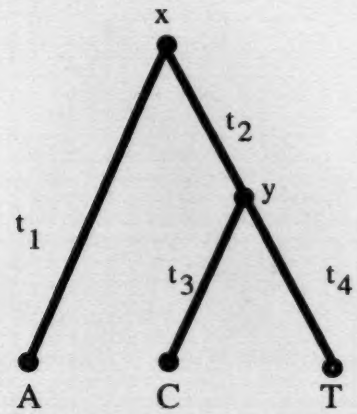


Figure 1.3: A phylogenetic tree, representing the data at a single site

The E_s and E_p are related to the free energy of the folded protein. The model reduces to the matrix 1.2.2 when $s = p = 0$. The parameter u is set for scaling the overall rate of change, while k is the transition/transversion bias and ω has the same interpretation as in the classic codon model 1.2.2.

The size of the matrix requires a different approach from the usual matrix exponentiation in order to calculate the transition probabilities from one protein sequence to another along a branch of the tree. Augmenting the data by creating a sequence of events between two sequences, ρ, an infinite number of possible paths can be imagined, but the problem can be formulated in Bayesian terms. Setting a prior density for the parameters $p(\theta)$ (i.e. branch lengths, elements of matrix 1.5.2), the joint posterior density is

$$p(\rho, \theta|i, j) = \frac{p(j, \rho|i, \theta)p(i|\theta)p(\theta)}{p(i, j)}.$$

Each path ρ between two sequences has defined a number of nucleotide substitutions with the intermediate states known and the time when each substitution occurred.

1.5.3 Ancestral state reconstruction

By using a set of extant protein coding sequences and the phylogeny relating them, one can predict nucleotide or amino acid sequences of their extinct ancestors. Knowledge of ancestral sequences for a set of extant sequences is very important for the study of evolutionary biology and also medical genetics. For example, an inferred ancestral sequence was used to design a HIV-1 vaccine [60]. Ancestral state reconstruction also furthered understanding of the evolution of SARS during the recent SARS epidemic

[61], the evolution of the steroid receptors [62] and mitochondrial metabolism [63], identified adaptive evolution in the bacteriolitic enzyme, lysozyme [64] and aided in reconstructing the visual pigments of dinosaurs [65].

The earliest ancestral state reconstruction methods were based on parsimony analysis, followed by methods that used stochastic models (maximum likelihood [66] [67] or Bayesian approaches [68] [69]) to find the most likely character states at the internal nodes. The parsimony method chooses the ancestral sequences that minimize the number of changes on the tree [53]. Parsimony makes no explicit assumptions about the pattern of amino acid substitution and the branch lengths have no relevance. The accuracy of ancestral states reconstruction is obviously low when sequence divergence is high. Yang *et al.* [67] combined branch length estimation using maximum likelihood with assignment of ancestral states using an empirical Bayes method. Yang implemented two variants of the method in the software package PAML [70]: marginal and joint reconstruction. The joint reconstruction method assigns ancestral states that give the maximum joint likelihood of the tree, while the marginal reconstruction finds the state at an internal node at a site that gives the maximum likelihood states at that site conditional on the tree. The review of Zhang and Nei (1997) [71] found that the maximum likelihood method outperformed the parsimony method in all the simulations in which sequence divergence was high, while the estimates were more similar when the sequence divergence was low. A full Bayesian approach was developed by Huelsenbeck and Bollback to accommodate the uncertainty of tree topology, branch lengths and parameters of the substitution model [69]. However, both model-based methods of phylogenetic inference commonly integrate over the four discrete states at the internal nodes, therefore without specifically generating the posterior distributions of character states. Because of this, inference of ancestral states is complicated and ambiguous. To overcome this drawback, Nielsen used data augmentation and a Bayesian framework to map the internal states onto the phylogeny treating them as variables in the chain [55]. Limiting the number of substitutions on branches to a maximum of two, a combination of Nielsen's method and the approach of Robinson *et al* [59] was developed by Krishnan *et al.* [72].

Ancestral sequence reconstruction involves a phylogenetic analysis to provide the most probable ancestral sequence. The protein obtained *in silico* is synthesized in the lab and further analyzed to detect potential structural and functional aspects not seen in the extant descendants.

The accuracy of ancestral protein reconstruction methods was assessed in many reviews ([73] [74] [75] [72]) with the general conclusion that the correctness of the inferred ancestral states is influenced by the tree topology, the branch lengths and the substitution model. However, to our knowledge, ancestral states reconstruction has not been evaluated in a model that allows continuous variation of substitution rates across sites.

In chapter 2 I develop a new method for Bayesian phylogenetic inference under very general substitution models using uniformization and data augmentation, that also allows the reconstruction of ancestral states.

1.6 Uncertainty in ancient DNA data due to miscoding lesions

Recent advances in molecular genetics allow DNA to be amplified and sequenced from ancient tissues [76] [77]. Conclusions drawn from a study of ancient DNA often generate a lot of interest in the scientific community, especially when they do not correspond to prior expectations. This is particularly true when human remains are analyzed. Recent criticisms focus on the possibility of contamination of the ancient samples with modern DNA. To eliminate this possibility, the DNA extraction procedures are very meticulous and the researchers have to follow a strict set of guidelines. However, the validation of an ancient sample as authentic encounters another difficult problem, post mortem damage. *In vivo* DNA damage is repaired by various enzymatic mechanisms, but once the metabolic pathways of cells have stopped, the DNA molecules start a progressive decay. The decay rate is influenced by a variety of factors related to the environment and the storage conditions. Biochemical processes subsequent to cell death cause the reduction of nucleotide sequence information in many ways (i.e. breakage of the DNA into 100-500 bp fragments, bases and sugar fragmentation, loss of amino groups) [78]. Some of the post mortem DNA modifications can block amplification during PCR, while others allow PCR products to be obtained, but incorrect bases might be incorporated and maintained in the amplification products. These kinds of PCR artifacts, commonly termed miscoding lesions, were first documented in 1989 by Pääbo [79] who observed the reduced amount of cytosine and thymine in the ancient samples. The biochemical explanation for the miscoding lesions consists in the hydrolytic deamination of cytosine to uracil and thymine and adenine to hypoxanthine and guanine. Because of strand complementarity and the difficulty of identifying the strand with original transition, the transitions A→G and T→C are termed as type I and C→T and G→A as type II, with type II being 30-50 times more frequent [80] in nuclear and mitochondrial genes [81]. The continuous improvement of amplification techniques reduced the number of such possible artifacts, but the exact rate or pattern of occurrence of miscoding lesions cannot be estimated. An approximate rate of post mortem damage was calculated by Hofreiter *et al.* (2001) [82] by comparing the PCR products of ancient samples with a database reference sequence. They concluded that miscoding lesion are unlikely to be more frequent than 0.1%. Because the working sequences are of a few hundred nucleotides, such a percent can cause faulty interpretations mimicking substitutions that cause evolutionary changes. An improved method to estimate the degradation rate was proposed by Gilbert *et al.*(2003) [83]. They define the degradation rate as the ratio of the hits observed at a specific site across all sequences analyzed to the total number of amplifications for that site. With regard to the distribution of damaged sites, the assumption of their random occurrence was found to be incorrect. Locations of several hotspots in mitochondrial DNA were identified in several studies [83] [84] [85]. Gilbert *et al.*(2003) [83] suggested a simple model to find such hotspots. They compared the expected and the observed number of substitutions under a Poisson process. However,

the degradation process is largely unknown. In addition, there is a need for new methodology to accommodate the uncertainty of the miscoding lesion in phylogenetic analysis and to investigate the evolutionary rates at degraded sites. As a correlation between post mortem and in vivo mutation rates has been found in a recent mitochondrial DNA study [84], it is important to have a comprehensive statistical method to analyze the possible role the degraded sites might have played.

In Chapter 4 I develop a discrete time Markov model of the process of post-mortem DNA damage in ancient samples and test the model using simulated data.

The main goal of my research is to analyze substitution rate variation across sites in a Bayesian formulation in a more realistic interpretation. The difficulty for allowing a continuous rate variation across sites (which is most likely the case based on empirical evidence) is given by the complexity of such a model. More specifically, the tedious calculations of nucleotide transition probabilities over time pose a real limitation on the complexity of DNA substitution models in phylogenetic inference. In other approaches, the problem is surpassed by using mathematical approximations, but which may bias the estimates of the parameters of interest: substitution rates, branch lengths or the parameters of the DNA substitution model. Understanding of the effects of such approximations is an important part of my research. Because there is no implementation which allows continuous rate variation across sites in phylogenetic inference for more than a few number of sequences, a significant part of my research was dedicated to write such a program.

The full Bayesian phylogeny analysis gives me the possibility to reconstruct the nucleotide sequences at the internal nodes in a statistically more realistic way then in many other available methods.

In the last part of my research, I extended the model by allowing ancient DNA (recent) in the analyzis. I am particularly interested in the errors that occur during chemical breakdown and how their presence might affect the inference of site specific rates.

Chapter 2

Estimation of site-specific rates

The theory presented in this chapter is published in *Systematics Biology* journal (L. Mateiu, B. Rannala, 2006, Inferring Complex DNA Substitution Processes on Phylogenies Using Uniformization and Data Augmentation Syst. Biol.55(2):259-269).

The calculation of transition probabilities presented in section 1.4 is required by all the parametric methods for phylogenetic inference that aim to infer the likelihood of a sample of DNA sequences. These transition probabilities have closed form solutions for simple models, such as the Jukes Cantor model [17], the Felsenstein model [14], etc, but more complex substitution models, such as the general time reversible model [16], do not have simple analytical solutions for the transition probabilities and these are instead calculated numerically by exponentiating the instantaneous rate matrix [58]. The computational expense of numerical substitution probability calculations via matrix exponentiation increases dramatically with an increase in the number of elements in the rate matrix. One of the major limitations on the complexity of the substitution models that may be used in phylogenetic inference is therefore the cost of numerically calculating the transition probabilities from the instantaneous rate matrix. Recently, Jensen and Pederson (2000) [86] have proposed a Markov chain Monte Carlo method for calculating transition probabilities for very complicated substitution models. The principle of the method is to model the complete set of (unobserved) nucleotide states visited by the chain along a branch separating two nodes. This allows arbitrarily complex models because only the transition probabilities for the states actually visited by the MCMC need to be specified. Such ideas have been applied to analyze complex substitution models with context-dependent rates of substitution; for example, sequences with overlapping reading frames under different selective pressures [87], models with dependent substitutions among codons determined by the structural properties of a protein [59], and models with dependent substitution rates among sites that account for such factors as CpG content [88].

The alternative approach I propose allows very complex substitution models to be used in phylogenetic inference and also makes use of MCMC methods to calculate transition probabilities. My approach does not require that the specific states that are visited be modeled, however, and instead only models the number of state changes on each

branch as an added variable in the MCMC. The method I propose relies on a "uniformization" (sometimes referred to as "randomization") of the Markov substitution process [56] [57]. The idea is quite simple, yet the resulting algorithm can potentially be much more efficient than calculating transition probabilities via matrix exponentiation, or augmenting the complete history of state changes, particularly in models that allow different sites to have different instantaneous rate matrices [89].

2.1 Model description

2.1.1 Data and parameters in the model

The model that I consider allows for rate variation across sites and branches. Let $\mathbf{x} = \{x_{kl}\}$ be a matrix of s aligned nucleotide sequences of n sites where x_{kl} is the nucleotide present at site l of sequence k. Let $\pi = \{\pi_{\mathtt{T}}, \pi_{\mathtt{C}}, \pi_{\mathtt{A}}, \pi_{\mathtt{G}}\}$ be the equilibrium nucleotide frequencies and let θ be the parameters of the substitution model. Let $\tau = \{T, \mathbf{w}\}$ represent an unrooted phylogeny of s species (e.g., a topology, T, and $2s-3$ branch lengths $\mathbf{w} = \{w_l\}$). Define $f(\mathbf{x}|\tau, \theta, \pi)$ to be the likelihood of the sequence data given the phylogenetic tree and other model parameters.

The focus of my research is to estimate $\mathbf{w}$ and other parameters of the substitution model considering the topology of the phylogenetic tree T as known. However, this approach can be extended to the inference of phylogenetic trees as well.

2.1.2 Uniformization of the Markov substitution process

Initially, I explain the method for calculating the substitution probabilities for a single site along a single branch of a tree. The GTR model allows each type of nucleotide substitution to have a separate rate, with the constraint that the process is reversible, so that for example the instantaneous rate of transition from A to C multiplied by the stationary frequency of A equals that from C to A multiplied by the stationary frequency of C, and so on. The instantaneous rate matrix of the GTR model, normalized so that the expected number of substitutions per unit time is 1, is

$$\mathbf{Q} = B \begin{pmatrix} -(a\pi_{\mathtt{C}} + b\pi_{\mathtt{A}} + c\pi_{\mathtt{G}}) & a\pi_{\mathtt{C}} & b\pi_{\mathtt{A}} & c\pi_{\mathtt{G}} \\ a\pi_{\mathtt{T}} & -(a\pi_{\mathtt{T}} + d\pi_{\mathtt{A}} + e\pi_{\mathtt{G}}) & d\pi_{\mathtt{A}} & e\pi_{\mathtt{G}} \\ b\pi_{\mathtt{T}} & d\pi_{\mathtt{C}} & -(b\pi_{\mathtt{T}} + d\pi_{\mathtt{C}} + \pi_{\mathtt{G}}) & \pi_{\mathtt{G}} \\ c\pi_{\mathtt{T}} & e\pi_{\mathtt{C}} & \pi_{\mathtt{A}} & -(c\pi_{\mathtt{T}} + e\pi_{\mathtt{C}} + \pi_{\mathtt{A}}) \end{pmatrix}$$

where the nucleotides are ordered T, C, A, G, the instantaneous rate matrix is multiplied by a normalizing constant [15],

$$B = \frac{1}{2}\left(\frac{1}{\pi_{\mathtt{T}}(a\pi_{\mathtt{C}} + b\pi_{\mathtt{A}} + c\pi_{\mathtt{G}}) + \pi_{\mathtt{C}}(d\pi_{\mathtt{A}} + e\pi_{\mathtt{G}}) + \pi_{\mathtt{G}}\pi_{\mathtt{A}}}\right),$$

and $a\pi_{\mathtt{T}}$ is the rate of substitution from nucleotide C to T, $b\pi_{\mathtt{A}}$ is the rate of substitution from nucleotide T to A, $\pi_{\mathtt{A}}$ is the stationary frequency of nucleotide A, etc. I use the

technique of uniformization [57] to transform the Markov process of DNA substitution into a time-homogeneous Poisson process in which substitution events occur with rate ν and the type of each substitution, conditional on a substitution event having occurred, is specified by a discrete Markov chain with probability elements

$$\mathbf{P} = \frac{B}{\nu}\begin{pmatrix} \nu(1/B - A_1) & a\pi_{\mathtt{C}} & b\pi_{\mathtt{A}} & c\pi_{\mathtt{G}} \\ a\pi_{\mathtt{T}} & \nu(1/B - A_2) & d\pi_{\mathtt{A}} & e\pi_{\mathtt{G}} \\ b\pi_{\mathtt{T}} & d\pi_{\mathtt{C}} & \nu(1/B - A_3) & \pi_{\mathtt{G}} \\ c\pi_{\mathtt{T}} & e\pi_{\mathtt{C}} & \pi_{\mathtt{A}} & \nu(1/B - A_4) \end{pmatrix},$$

where $\nu = 1/\pi_{\min}$ and $\pi_{\min} = \min_i \pi_i$, for all $i \in \{\mathtt{G}, \mathtt{C}, \mathtt{A}, \mathtt{T}\}$ is the smallest nucleotide frequency. For the normalized instantaneous rate matrix, $\sum_{i \neq j} \pi_i Q_{ij} = 1$ and therefore $\pi_i Q_{ij} \leq 1$ and $Q_{ij} \leq 1/\pi_i$ so that $1/\pi_{\min}$ is a bound on the maximum rate. The empirical nucleotide frequencies in the sampled sequences are used as estimates of the stationary nucleotide frequencies. I define,

$$\begin{aligned} A_1 &= \frac{1}{\nu}(a\pi_{\mathtt{C}} + b\pi_{\mathtt{A}} + c\pi_{\mathtt{G}}), \\ A_2 &= \frac{1}{\nu}(a\pi_{\mathtt{T}} + d\pi_{\mathtt{A}} + e\pi_{\mathtt{G}}), \\ A_3 &= \frac{1}{\nu}(b\pi_{\mathtt{T}} + d\pi_{\mathtt{C}} + \pi_{\mathtt{G}}), \\ A_4 &= \frac{1}{\nu}(c\pi_{\mathtt{T}} + e\pi_{\mathtt{C}} + \pi_{\mathtt{A}}). \end{aligned}$$

The probability that a substitution from nucleotide i to j occurs on a branch of length w, $p_{ij}(w)$, can then be written as the infinite sum

$$p_{ij}(w) = \sum_{M=0}^{\infty} \frac{(\nu w)^M e^{-\nu w}}{M!} \times P_{ij}^{(M)}, \tag{2.1}$$

where $P_{ij}^{(M)}$ denotes element i, j of the Markov chain derived for the discretized process under uniformization raised to the Mth power.

The sum over the number of transitions (eq. 2.1) is evaluated with the use of MCMC.

2.1.3 Modeling Rate Variation Among Sites

To illustrate the method I apply the algorithm to estimate branch lengths and site-specific substitution rates assuming a continuous gamma distribution as the prior for rates across sites. Let $\mathbf{r} = \{r_m\}$ be a vector of site specific rates (of length n), where r_m is the rate for site m. Define $f(r_m|\alpha)$ to be the prior density of rates for the mth site with α to be the parameters of the prior on rates. The marginal posterior probability of the phylogeny can be obtained by taking the expectation over the prior density of site-specific rates [15],

$$f(\tau|\theta, \mathbf{x}, \alpha, \lambda, \pi) = \mathtt{C}(\theta, \pi, \alpha, \lambda, \mathbf{x}) f(\tau|\lambda) \prod_{m=1}^{n} \mathtt{E}\left[f(\mathbf{x}_m|\tau, r_m, \theta, \pi) f(r_m|\alpha)\right], \tag{2.2}$$

where $f(\tau|\lambda)$ is the prior on phylogenetic trees and $\mathtt{C}(\theta, \pi, \alpha, \lambda, \mathbf{x})$ is a normalizing constant obtained by integrating the equation to the right of C over all tree topologies and branch lengths,

$$\frac{1}{\mathtt{C}(\theta, \pi, \alpha, \lambda, \mathbf{x})} = \int_{\tau} f(\tau|\lambda) \prod_{m=1}^{n} \mathtt{E}\left[f(\mathbf{x}_m|\tau, r_m, \theta, \pi) f(r_m|\alpha)\right] d\tau.$$

If one is primarily interested in estimating site-specific rates and substitution model parameters, rather than phylogeny, the problem can be reformulated as

$$f(\mathbf{r}, \theta|\mathbf{x}, \alpha, \lambda, \pi) = \mathtt{C}(\pi, \alpha, \lambda, \mathbf{x}) \int_{\tau} f(\tau|\lambda) \prod_{m=1}^{n} f(\mathbf{x}_m|\tau, r_m, \theta) f(r_m|\alpha) f(\theta) d\tau. \quad (2.3)$$

If the tree topology is known, the integral is evaluated over the branch lengths, otherwise it is an integral over the branch lengths and a sum over the topologies. Similarly, the joint probability density of site-specific rates, substitution model parameters, and branch lengths conditioned on topology, T, is

$$f(\mathbf{r}, \mathbf{w}, \theta, \alpha, \lambda|\mathbf{x}, \pi, T) = \mathtt{C}(\pi, T, \mathbf{x}) f(\mathbf{w}|\lambda) \prod_{m=1}^{n} f(\mathbf{x}_m|\tau, r_m, \theta) f(r_m|\alpha) f(\theta) f(\alpha) f(\lambda). \quad (2.4)$$

The focus of this study is to evaluate the probability density presented in equation 2.4 above.

Augmented likelihood

I use data augmentation to integrate over two additional vectors of random variables, the numbers of transitions on each branch and the unobserved ancestral nucleotides at the internal nodes of the tree. Define $\mathbf{M} = \{M_{lm}\}$, where M_{lm} is the number of transitions at site m on branch l of a phylogenetic tree T. Further, let $\mathbf{x}^- = \{x^-_{kl}\}$ be a matrix of the $s-2$ ancestral nucleotide sequences on the tree. Define $\theta = \{a, b, c, d, e\}$ to be a matrix of the parameters of the GTR substitution model (with a 5RR parametrization; [90]). The augmented likelihood is

$$f(\mathbf{M}, \mathbf{x}, \mathbf{x}^-|\mathbf{r}, \tau, \pi, \theta) = \prod_{m=1}^{n} \prod_{l=1}^{2s-3} f(\mathbf{x}_m, \mathbf{x}^-_m|\theta, M_{lm}, r_m, w_l, \pi, T) \Pr(M_{lm}|r_m, w_l). \quad (2.5)$$

According to the theory developed above, the probability of M_{lm} transitions at site m on branch l in the uniformized Markov process is Poisson with probability distribution

$$\Pr(M_{lm}|r_m, w_l) = \frac{e^{-\nu w_l r_m} (\nu w_l r_m)^{M_{lm}}}{M_{lm}!}.$$

The probability of a change from nucleotide i to j at site m on branch l, given M_{lm} transitions, is $\mathtt{P}_{ij}^{(M_{lm})}$ (this is the conditional likelihood).

It is also possible to explicitly sum over the ancestral nucleotides using the usual pruning algorithm [14] to calculate the likelihood conditional on the number of transitions.

Posterior probability density of rates and branch lengths

Following Yang (1993) [15] the site-specific substitution rate parameter is assumed to have a prior density that is a gamma distribution with mean one and shape parameter α so that

$$f(r_m|\alpha) = \frac{\alpha e^{-\alpha r_m}(\alpha r_m)^{\alpha-1}}{\Gamma(\alpha)}.$$

I assume an exponential distribution with common parameter λ for w_l and I use the Dirichlet prior for θ suggested by Zwickl and Holder (2004) [90]. I use uniform hyperpriors on λ and α and I use empirical estimates for π. The posterior density is then,

$$f(\mathbf{r},\mathbf{w},\theta,\alpha,\lambda|\mathbf{x},\pi,T) = \sum_{\mathbf{M}}\sum_{\mathbf{x}^-} f(\mathbf{w}|\lambda)f(\mathbf{M},\mathbf{x},\mathbf{x}^-|\mathbf{r},\tau,\pi,\theta)f(\mathbf{r}|\alpha)f(\theta)f(\alpha)f(\lambda). \tag{2.6}$$

The density of equation 2.6 is evaluated using MCMC.

2.1.4 Tree likelihood calculation using uniformization and data augmentation in the Bayesian approach

The general Bayesian formula to be evaluated in a MCMC with Metropolis Hastings algorithm contains the likelihood ratio multiplied by the prior ratio and the proposal ratio. If the proposal of the new value for a parameter in the equation is symmetrical (e.g., the proposal probability of state i given j is equal to that of j given i), the proposal ratio is equal to one. If the prior for a parameter is uniform, the prior ratio is also equal to one. The beauty of the Bayesian method coupled with MCMC and Metropolis Hasting algorithm [91] is that only the ratio of terms involving the newly proposed and the current value have to be calculated as all other terms do not change and, therefore, parameter value cancels out in the ratio. Following the uniformization idea, the substitution rate parameter at each site is separated from the transition probability matrix and the substitution events on the branches of the tree are random variables in the Markov chain. The transition probability matrix, containing the nucleotide substitution probabilities along a branch of the tree, is related only to the number of the substitutions on the branch. As the one-step matrix is shared by all the sites and branches, the calculations of the tree likelihood are greatly simplified, allowing us to augment the data and treat nucleotides at the internal states as random variables in the Markov chain.

At every step in the chain, one-by-one, and in random order, every branch, the length, site rate and substitution number along the branch are updated. The nucleotide at an end of the branch is also updated by simply proposing and evaluating the new nucleotide without the need to extend the summation to the subtree above the previously chosen

branch. In other words, the nucleotides at the internal states become random variables in the MCMC. This can be more easily explained by referring to figure 1.3. An internal node is randomly picked, the root node x with the nucleotide T for instance. The nucleotide is proposed to be changed by picking a random nucleotide, giving equal chance to any of the four to be chosen. Let us assume the nucleotide A is picked. At this stage, the nucleotide at node y is fixed at C. The new tree likelihood is compared with the old one in the Metropolis Hastings ratio:

$$\frac{Prob_{new}(A,C,T|\tau)}{Prob_{old}(A,C,T|\tau)} = \frac{Prob(A)}{Prob(T)} \frac{Prob(A|T,t_1)}{Prob(A|A,t_1)} \frac{Prob(C|A,t_2)}{Prob(C|T,t_2)}$$

This simple example shows clearly the difference between calculations using conventional methods and the approach I am pursuing. Instead of evaluating all the summations in equation 1.3, just the product of the three terms has to be evaluated.

Computational complexity

Equation (2.4) can, in principle, be evaluated directly via MCMC methods. However, it is clearly computationally expensive to do so for non-trivial substitution models. For example, each time a new site-specific rate is proposed in the MCMC one must recalculate transition probabilities for each of the $(2s-3)$ branches. If one diagonalizes the rate matrix (to allow exponentiation of the rate matrix to calculate the transition probabilities), a calculation of the marginal likelihood for one branch (applying the pruning algorithm) requires $2h^2+h$ operations (where h is the dimension of the substitution matrix). This ignores the initial cost of calculating the eigenvalues and eigenvectors of the rate matrix, which only needs to be done once if the MCMC is integrating only over r and w. In the uniformized MCMC calculation, the log of the Metropolis-Hastings ratio (when a rate change is proposed for a site) is a simple difference of proposed and current rates and of the logs of proposed and current rates, multiplied by the number of transitions, for each branch.

2.2 Implementation

A computer program was written in $C/C++$ to process the data, implement the theory above and provide various statistics during and after the run (approximately 4000 lines). I called the program BYPASSR (BaYesian Phylogenetic Analysis of Site Specific Rates). Part of the program manual is attached in Appendix I.

Several DNA substitution models are implemented in BYPASSR: GTR [16], TN93 [92], HKY85 [40], F84 [93] and JC69. Their parameterization follows the PAML [94] implementation. The rate matrices for these models are:

$$GTR : \mathbf{Q} = \begin{pmatrix} * & a\pi_C & b\pi_A & c\pi_G \\ a\pi_T & * & d\pi_A & e\pi_G \\ b\pi_T & d\pi_C & * & \pi_G \\ c\pi_T & e\pi_C & \pi_A & * \end{pmatrix} \tag{2.7}$$

$$TN93 : \mathbf{Q} = \begin{pmatrix} * & k_1\pi_C & \pi_A & \pi_G \\ k_1\pi_T & * & \pi_A & \pi_G \\ \pi_T & \pi_C & * & k_2\pi_G \\ \pi_T & \pi_C & k_2\pi_A & * \end{pmatrix}, \tag{2.8}$$

$$F84 : \mathbf{Q} = \begin{pmatrix} * & (1+k/\pi_Y)\pi_C & \pi_A & \pi_G \\ (1+k/\pi_Y)\pi_T & * & \pi_A & \pi_G \\ \pi_T & \pi_C & * & (1+k/\pi_R)/\pi_G \\ \pi_T & \pi_C & (1+k/\pi_R)\pi_A & * \end{pmatrix} \tag{2.9}$$

$$HKY85 : \mathbf{Q} = \begin{pmatrix} * & k\pi_C & \pi_A & \pi_G \\ k\pi_T & * & \pi_A & \pi_G \\ \pi_T & \pi_C & * & k\pi_G \\ \pi_T & \pi_C & k\pi_A & * \end{pmatrix}, \tag{2.10}$$

$$JC : \mathbf{Q} = \begin{pmatrix} * & 1 & 1 & 1 \\ 1 & * & 1 & 1 \\ 1 & 1 & * & 1 \\ 1 & 1 & 1 & * \end{pmatrix} \tag{2.11}$$

The $*$ symbols on diagonals correspond to negative values such that each line sums to 0, π_Y and π_R are the sums of pyrimidine and purine frequencies, respectively and k is the transition/transversion bias. After a phylogenetic tree and a DNA substitution model are selected, Equation 2.1 is applied to calculate the marginal probability of the parameters in the table 2.1.

2.2.1 Metropolis Hastings algorithm

Modifying the number of substitution events at a branch and a site

To formulate the problem in terms of an MCMC algorithm, note that Equation 2.1 can be written as a marginal probability for the transition from nucleotide i at one end of a branch to j at the other end, with the expectation taken over M,

$$p_{ij}(w) = \sum_{M=0}^{\infty} \Pr(M, i \to j)$$

	parameter	prior distribution for parameter
r	site specific rates	$Gamma(\alpha, \alpha^2/\beta)$ Yang (1993) [15]
w	branch lengths	$Exponential(\lambda)$ $Uni(0, 20)$
x^-	{A,T,C,G}	$DiscreteUniform(0, 4)$
M	unif. substitutions events	$Uniform(0, 50)$
θ	param. GTR subst.model	$EquivalentDir(1, 1, 1, 1, 1)$ Zwickl and Holder (2004) [90]
α	param. of Gamma Distr.	$Uniform(0, 100)$
λ	param. of Exp Distr.	$Uniform(0, 100)$

Table 2.1: Parameters and their prior distribution in the model.

and the Metropolis-Hastings algorithm can then be used to obtain the marginal distribution, rather than evaluating the sum explicitly. One simple implementation is to use a symmetrical proposal density for M: $g(M^*) = 1/3$ if $M^* = M$, $M^* = M - 1$ or $M^* = M + 1$ and $M \neq 0$ or $g(M^*) = 1/3$ if $M = 0$ and $M^* = 1$, or $g(M^*) = 2/3$ if $M = 0$ and $M^* = 0$. An initial value for M is randomly assigned from the positive integers and at each iteration of the algorithm a new state M^* is proposed for M from $g(.)$ and accepted with probability

$$\alpha = \min\left\{1, \frac{e^{-\nu w}(\nu w)^{M^*}/M^*!P_{ij}^{(M^*)}}{e^{-\nu w}(\nu w)^{M}/M!P_{ij}^{(M)}}.\right\}$$

For $M > 1$, the ratio at the right of the above equation simplifies to become $(P_{ij}^{(M+1)}/P_{ij}^{(M)})\times$ $\nu w/(M+1)$ if $M^* = M + 1$ and $(P_{ij}^{(M-1)}/P_{ij}^{(M)}) \times M/(\nu w)$ if $M^* = M - 1$.

The formula differs if the pruning algorithm is used instead,

$$\alpha = \min\left\{1, \frac{e^{-\nu w}(\nu w)^{M^*}/M^*!P^{(M^*)}}{e^{-\nu w}(\nu w)^{M}/M!P^{(M)}},\right\}$$

where P^M is the marginal probability of the nucleotides of the sampled sequences, obtained by summing over all ancestral states conditional on the number of changes on the branch.

Calculating the transition probabilities as outlined above has the advantage of allowing one to integrate out over M in a MCMC analysis, augmenting the data by treating M as an unobserved random variable in the chain. This is particularly useful for implementing site-specific rates because the substitution rate parameter r only occurs as a simple term in the Metropolis-Hastings ratio and does not feature in the discrete Markov chain determining the conditional substitution probabilities. Note that $rt = w$, where t

is the branch length in units of time (using the same timescale as was used to specify the rate r) whereas w is the branch length in units of expected numbers of substitutions. This allows a common substitution matrix to be applied across sites with only a simple recalculation of the weighting term across branches when a new rate is proposed for a specific site. The trade-off is that the MCMC algorithm must now also integrate over the numbers of substitutions on each branch. The number of substitutions follow a Poisson process with rate $\nu r_i t$, where r_i is the substitution rate at site i. The uniformization constant calculated as the inverse of the smallest nucleotide frequency can occasionally take values of 10 or more which, when multiplied with a high rate of 10, 15 or even 30 at a fast evolving site and a branch length close to 1, can result in an expected number of substitutions under uniformized process that exceeds 100. However, for most datasets, the expected number of substitutions per branch under uniformized process is small (usually less than about 20), so a relatively low number of matrix powers are needed. The properties of stochastic matrices (matrices with sum on each line equal to 1) can be used to speed up the calculations. A stochastic matrix has an asymptotic convergence to the stationary distribution [95] (in my case the base frequencies) and a number N exists such that

$$|P_{ij}^{\prime(N)} - P_{ij}^{\prime(N-1)}| < \epsilon, \tag{2.12}$$

for all i, j where ϵ is a desired accuracy. In other words, if the absolute difference between each pair of elements of the matrix raised at two consecutive powers is less than a specified error ϵ (e.g. 0.005), the matrices raised at powers greater than N has approximately the same values as the matrix power N.

Also, because the transition matrix for the discrete process does not depend on the substitution rate parameter, this matrix calculation only needs to be performed once if one is integrating over the substitution rates alone.

Modifying branch length

In Bayesian phylogenetics, exponential and uniform densities are the usual priors for branch lengths (Fig 2.1). The uniform prior assigns values from 0 to an arbitrary (i.e. 20) upper bound with equal probabilities. In general, the branch lengths have subunitary values and there is no need to investigate such a large sampling space. It is not computationally efficient [96]. The exponential prior is a continuous probability density that goes from 0 to infinity. An upper bound is also set, but it has only theoretical meaning because such extreme values have very low probabilities and, in practice, are almost never visited. In general, the exponential prior is preferred because of the reduced proportion of large biologically unreasonable values. I have implemented both priors. The results of simulations with both are shown in the following section.

A branch i is chosen at random. The candidate branch length is chosen from the interval $(0, 20)$ altering the current value with a random number, using a sliding window. If the current branch length is w and the tuning parameter is δ, then I generate a uniform random number u on the interval $(-\delta,+\delta)$ and propose $w' = w + u$. Because all the

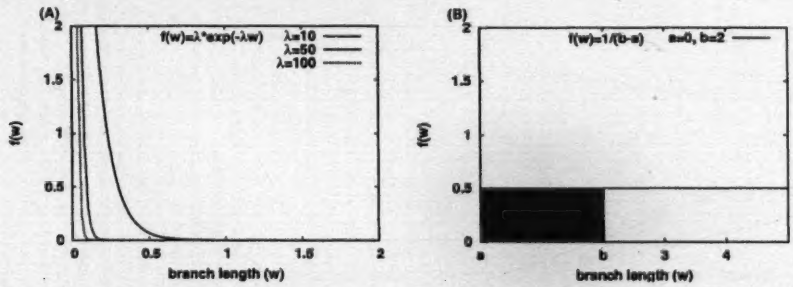


Figure 2.1: Exponential (A) and uniform (B) densities of branch lengths prior.

variables in the chain must have positive values, if a proposed u makes the branch length negative, the absolute value of w' is taken instead. If the proposed value w' exceeds the upper bound U of the interval, then $w' = 2 * U - w'$.

The new branch length w' is accepted with a probability

$$R = \left\{1, \left[\left(\prod_{j=1}^{N} e^{-r_i \nu (w'-w)} \left(\frac{w'}{w}\right)^{M_{ij}}\right) e^{-\lambda(w'-w)}\right]\right\} \tag{2.13}$$

where M_{ij} are the substitutions at site j for the chosen branch i.

If the uniform prior is used instead, the term outside the product (prior on w') is dropped.

After updating the branch length, the number of substitutions on that branch for all the sites is also updated as detailed previously. The nucleotides are also updated for all the sites at the nodes adjacent to the branch i. If branch i has the root as its descendant, a new nucleotide at the root is chosen at random (e.g. assuming uniform probability for each nucleotide). The move is evaluated using the corresponding terms from the uniformized substitution matrix given the nucleotides at the descendant nodes of the root. The Metropolis Hastings ratio is

$$R = \left\{1, \left[\frac{\pi_{x'} P^{(M)}_{x'y} P^{(M)}_{x'z}}{\pi_x P^{(M)}_{xy} P^{(M)}_{xz}}\right]\right\}, \tag{2.14}$$

where y is the nucleotide at the left descendant node of the root, z is the nucleotide at the right descendant node of the root and π is the base frequency at the root node.

If the branch is not adjacent to the root, the equation becomes:

$$R = \left\{1, \left[\frac{P^{(M)}_{x'y} P^{(M)}_{x'z} P^{(M)}_{x'u}}{P^{(M)}_{xy} P^{(M)}_{xz} P^{(M)}_{xu}}\right]\right\}, \tag{2.15}$$

where y is the nucleotide at the left descendant node, z is the nucleotide at the right descendant node and u is the nucleotide at the parent node.

I observed a faster convergence of the chain by choosing this approach than updating the nucleotides at a random node or the number of transitions on a random branch.

Modifying rate at a site

Next a site j is randomly chosen. The modified rate is reflected if it falls outside the defined bounds. The lower bound is 0, while the upper bound is chosen to be 100. The proposed rate r' is accepted with the probability:

$$R = \left\{1, \left[\left(\prod_{i=1}^{2s-3} e^{-\nu t_i(r'-r)} \left(\frac{r'}{r}\right)^{M_{ij}}\right) e^{-\alpha(r'-r)} \left(\frac{r'}{r}\right)^{(\alpha-1)}\right]\right\} \tag{2.16}$$

New numbers of substitutions are proposed for all the branches at that site. The nucleotides for all the nodes at the chosen site are also updated. The procedure is the same as for the branch length move.

Modifying α

The shape parameter α has an upper bound of 100. A new value of α is proposed using a sliding window and is accepted according to the ratio:

$$R = \left\{1, \prod_{j=1}^{N} \left(e^{-r_j(\alpha'-\alpha)} r_j^{\alpha'-\alpha} \frac{\alpha'^{\alpha'}}{\alpha^{\alpha}} \frac{\Gamma(\alpha)}{\Gamma(\alpha')}\right)\right\} \tag{2.17}$$

Modifying λ

The parameter of the exponential prior distribution on branch lengths, λ, has an upper bound of 500. At each cycle a new value for λ is proposed using a sliding window and the proposed value of λ is accepted with probability:

$$R = \left\{1, \prod_{i=1}^{2s-3} \frac{\lambda'}{\lambda} e^{-t_i(\lambda'-\lambda)}\right\}. \tag{2.18}$$

Modifying the rate parameters of the DNA substitution model

GTR

I use the equivalent of the uniform Dirichlet prior as described in Zwickl *et al.*(2004) [90] for the relative rates of the substitution matrix, a, b, c, d, e. For a parameterization of substitution rate matrix with f set to 1 (so called, 5RR parameterization), the alternate for uniform Dirichlet prior is

$$P(a,b,c,d,e) = \frac{1}{5!(1+a+b+c+d+e)^6} \tag{2.19}$$

The relative rate parameters in BYPASSR are updated all at once every fourth iteration. The algorithm is as follow:

- $S_1 = a + b + c + d + e$

- $a' = a \pm \delta_a$; $b' = b \pm \delta_b$; $c' = c \pm \delta_c$; $d' = d \pm \delta_d$; $e' = e \pm \delta_e$ with $a', b', c', d', e' \in (0, 50)$

- $S_2 = a' + b' + c' + d' + e'$, where the deltas are the tuning parameters for the relative rates.

Then, I calculate the acceptance probability for all 5 rates at once.

$$R = \left\{1, \left[\prod_{i=0}^{branches} \prod_{j=0}^{sites} \frac{(P')_{ij}^{M}}{(P)_{ij}^{M}}\right] \left(\frac{1+S_2}{1+S_1}\right)^6\right\}$$

After the relative rates are updated, a vector of the uniformized matrices shared by all the sites and branches is calculated iteratively and stored for all the powers given by the maximum number of substitutions allowed, but truncated when the difference between two consecutive powers is less than an error value ϵ as given in Eq 2.12.

F84, HKY85, TN93

The K from F84 and HKY85, K_1, K_2 are modified in the same way as any of the parameters in the GTR, but the prior term is missing from the acceptance probability equation.

The sampling algorithm can be summarized in this way:

```
for nrnodes/nrsites
    pick a random branch
      move branch length
      move nr transitions at branch node sites
      move nucl at branch node sites

for nrsites/nrnodes
    pick a random site
      move rate
      move nr transition at site on branches
      move nucl at site on branches

move alpha

move lambda
```

The two loops ensure an approximately equal number of passes through branches and sites. It is addressed to datasets with few sequences and large number of sites or reversed.

2.3 Statistical performance

2.3.1 Simulating data to evaluate the method

Parametric methods for phylogenetic reconstruction are based on a clearly defined DNA substitution model allowing genetic data to be simulated under the model for a specified set of parameters. Analysis of simulated data is a very important step in validating a new method. The statistical performance of the method I proposed was tested on nucleotide sequences generated by two programs. The PAML [94] package includes the EVOLVER, which uses Monte Carlo simulation to generate a set of nucleotide sequences. The input parameters are: a phylogenetic tree with specified branch lengths, a DNA substitution model with the desired parameters, including the nucleotide frequencies and the shape parameter of a continuous gamma distribution as the prior distribution for site specific rates.

I have also written a second program (BYPASSR-gen), to generate a random topology with branch lengths from an exponential prior, or with ultrametric branch lengths as expected under the molecular clock. Simulation of the nucleotide substitution process follows the uniformization idea. The number of substitutions on each branch is simulated under a Poisson process with rate $\nu r_i t$, where r_i is the substitution rate at site i. The nucleotide at a node and site is chosen according to its probability under the uniformized substitution matrix raised to the power corresponding to the number of substitutions on the branch adjacent to the node. The other parameters to be specified are the α parameter of the gamma distribution and the nucleotide frequencies. Both methods gave identical results for the simulation study, so I will not make further distinction as to the method used.

2.3.2 Comparison with maximum likelihood implementation in PAML

The PAML [70] package contains a set of programs for phylogenetic analysis using the maximum likelihood method. One of the modules BASEMLG assumes a continuous gamma density for the site-specific rates. The mean of the conditional distribution of site-specific rates from BASEMLG is directly comparable with the mean of the posterior densities of rates from BYPASSR. The use of BASEMLG is restricted to a small number of sequences (e.g. 7, 8) and only a few simple DNA substitution models are practicable because of computational limitations. BASEML uses a discrete approximation for the gamma distributed prior on site-specific rates of DNA sequences, allowing more sequences to be used, but the adequacy of the discrete category approximation is not well known.

CODEML is the one of the programs most commonly used to detect selection at individual protein-coding sites. Based on a codon substitution model [29], CODEML estimates the proportion of sites that have dN/dS ratio greater, smaller or equal to one, confirms the significance of positive selection using likelihood ratio tests and identifies specific codon sites under positive selection using empirical Bayesian methods. The

review of Yang and Bielawski (2000) [97] provides an extensive list of genes where CODEML successfully identified positive selection operating along codon sites [97]. Positive selection was found in many viral genes as well as mammalian genes involved in immunity, reproduction, digestion etc.

BYPASSR vs. BASEMLG

The program BASEMLG uses as its point estimate of the site specific rate the conditional expectation of the site-specific rates obtained by integrating over the conditional distribution of rates with other model parameters fixed at their maximum likelihood values [98]. This is an empirical Bayes estimator. Using datasets of 6 or 7 taxa generated and analyzed with simple DNA substitution models (Hasegawa Kishino Yano [99] and Jukes Cantor [17], respectively), I compared the estimates of the two programs with a continuous gamma implementation for site specific rates. The relationship between the estimates is very consistent and linear, suggesting that the two methods are producing similar rate and branch lengths estimates, as expected. Fig. 2.2 shows the plot of site rate estimates obtained with the two methods for a dataset of 6 sequences and 1000 sites. The amount of information about site-specific rates is mainly determined by the number of sequences sampled and the reduced amount of information contained in such a small number of sequences, therefore generates large variances in the rate estimates. This explains a correlation coefficient of only 0.749. However, the relationship between the rate estimates of the two methods is improved by adding another sequence. The results for the analysis of 7 taxa and 1000 sites dataset are shown in Fig. 2.3).

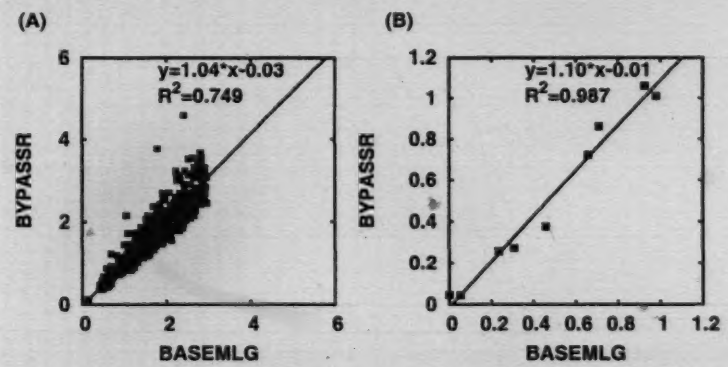


Figure 2.2: Comparison of estimated site-specific rates (panel A) and branch lengths (panel B) obtained using BYPASSR and BASEMLG programs for 6 taxa and 1000 sites dataset generated and analyzed assuming KHY85+Γ model.

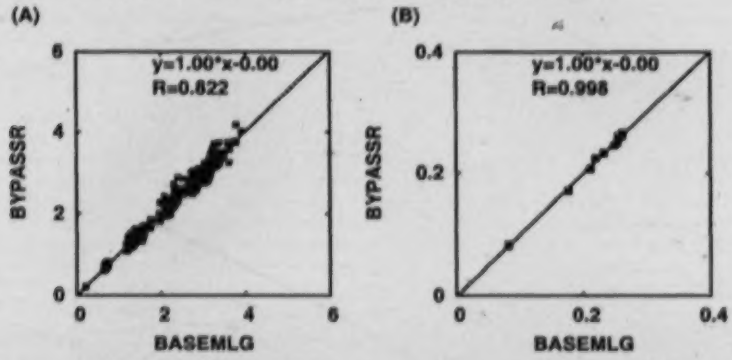


Figure 2.3: Comparison of estimated site-specific rates (panel A) and branch lengths (panel B) obtained using BYPASSR and BASEMLG programs for 7 taxa and 1000 sites dataset generated and analyzed assuming JC+Γ model.

BYPASSR vs. BASEML

A larger simulation study was carried out to examine the effects of both sequence length, $n \in (500, 2000)$, and number of taxa, $s \in (10, 20, 50, 250)$ on the accuracy of site-specific rate inferences. For each of several combinations, two datasets were simulated using the following procedure: (1) generate a random tree from a birth-death process (all labeled histories equally likely); (2) simulate branch lengths from an exponential prior with $\lambda = 20$; (3) Simulate site-specific rates using a gamma distribution with $\alpha = 0.5$ and (4) Simulate sequences under a GTR model with parameters $a = 0.25$, $b = 0.75$, $c = 1.25$, $d = 1.75$, $e = 2.25$ $\pi_T = 0.1$, $\pi_C = 0.2$, $\pi_A = 0.3$, $\pi_G = 0.4$. The effect of using a discrete gamma approximation [50] on the accuracy of estimates of site-specific rates obtained using BASEML was examined for various numbers of rate categories. The BASEML program offers two options for obtaining point estimates of site-specific rates. The first option uses a weighted average of the rate for each category multiplied by the conditional probability of the category. This is a discrete approximation to the conditional expectation used in BASEMLG. The second option uses the rate for the site class having the highest posterior probability. I used the first option in the analysis.

Figures 2.4 and 2.5 are typical of the results obtained. Here the mean posterior rate (from BYPASSR) and the weighted mean of the rate for each site (from BASEML) are plotted against the actual rate for each site (Figures 2.4 and 2.5, respectively). It is evident that even with 20 rate categories, the rate estimates obtained using the discrete approximation tend to underestimate the true rates and this is most evident with 5000 sites because more extreme rates are observed when more sites are examined.

Another interesting observation from the simulation study is that increasing the number of sites has little effect on the variance of the posterior distribution of rates from BYPASSR while increasing the number of taxa has a very dramatic effect (Figure 2.6). Panel A of Figure 2.6 shows the posterior distributions obtained for a site with true substitution rate $r = 0.35$, with either 10, 20, 50 or 250 taxa and $n = 500$ sites sampled.

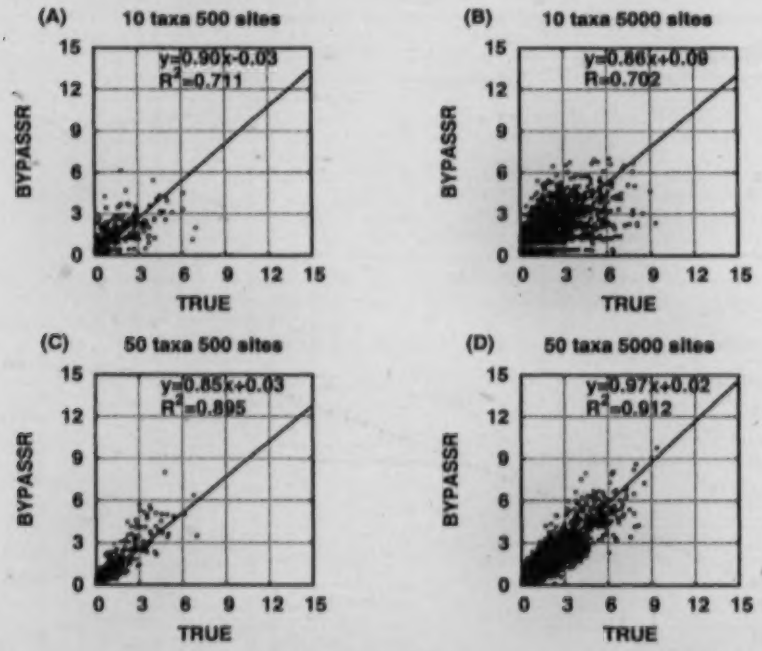


Figure 2.4: Plot of mean posterior site-specific rates from the BYPASSR program versus true rates. Data for the plot were simulated using either 10 taxa (panels A and B) or 50 taxa (panels C and D) and either 500 sites (panels A and C) or 5000 sites (panels B and D) under a GTR model.

With only 10 taxa the posterior looks essentially identical to the prior (in this case, a gamma distribution with $\alpha = 0.5$). With increasing numbers of taxa, however, the distribution becomes more modal with the mode shifting towards the location of the true rate. Panel B of Figure 2.6 shows the results for another simulation with the true rate at a site to be $r = 0.35$ and $n = 5000$ sites. In this case, the posterior densities for 10, 20 and 50 taxa are very similar to those observed in panel A ($n = 500$ sites). In general, the posterior density is much more concentrated with a clear mode when rates are in the intermediate range. Panels C and D of Figure 2.6 show the posterior densities obtained using 10, 20, 50 or 250 taxa with either a much higher rate ($r = 2.17$) (panel C of Figure 2.6) or a much lower rate ($r = 0.09$)(panel D of Figure 2.6). In both cases, the variance of the posterior is increased and estimates are clearly influenced by the prior for fewer than 250 taxa. Because the mean rate in the prior is 1, estimates based on a small number of taxa for sites with very low rates tend to have positive bias (overestimating true rate) and for sites with very high rates tend to have negative bias (underestimating true rate). Clearly, a large number of taxa are needed to get precise estimates of site-specific rates.

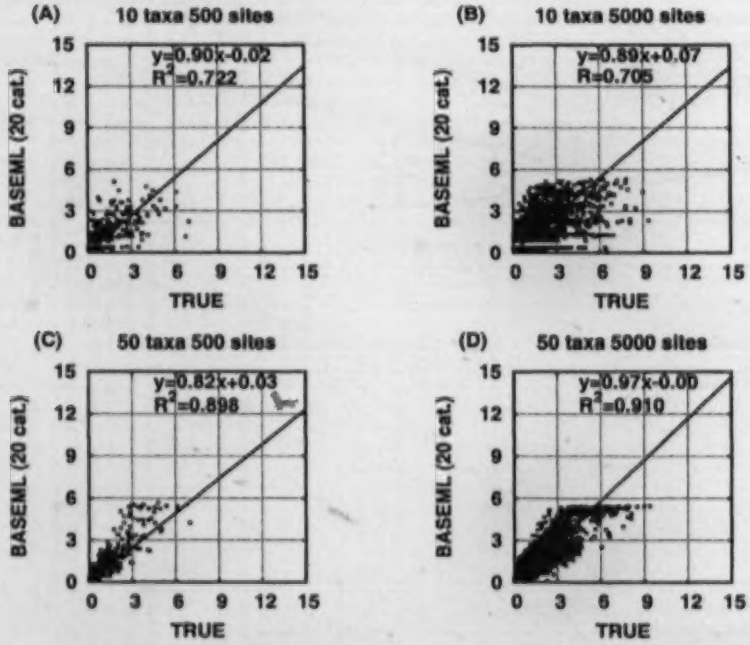


Figure 2.5: Plot of empirical Bayes estimates of rates obtained using the BASEML program (with 20 rate categories for the discrete gamma approximation) versus true rates. Data for the plot were simulated using either 10 taxa (panels A and B) or 50 taxa (panels C and D) and either 500 sites (panels A and C) or 5000 sites (panels B and D) under a GTR model.

2.3.3 The influence of topology on substitution rate estimates

BYPASSR assumes no recombination within genes. If recombination occurs, the topology for the recombinant sites differs from the remaining sites. A simple way to test the performance of my method when the assumption of no recombination does not hold is to analyze the dataset using random trees. In this case, none of the sites has a correct tree [100]. The interconnection between the gamma-distribution shape parameter α and tree topology was previously examined by Yang *et al.* [50] [9] using DNA substitution models. They concluded that reliable estimates of α are obtained even if the topology was not correct. I investigate the rate distribution at each site when wrong tree topologies are used in a limited number of runs on simulated data. Because the number of sites has little effect on site specific rate estimates, I generated datasets of different numbers of sequences 25, 50, 75 and 1000 sites. Substitution rates were simulated from a gamma distribution with either $\alpha = 0.1$, 0.5 or 1. The same site rates were used for all datasets with a given α and a common topology. The results are summarized in Table 2.2. The correlation coefficient shows a weaker positive correlation when the tree topology was

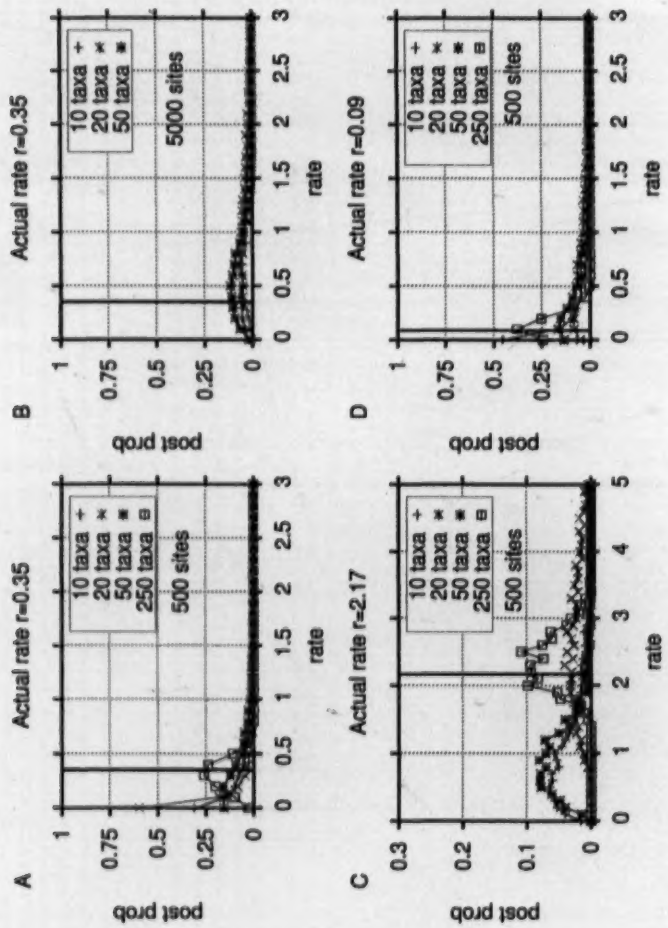


Figure 2.6: Plot of posterior distribution of site-specific rates for simulated data analyzed using the BYPASSR program. Panels A and B show posterior distributions for different numbers of taxa with an actual rate of 0.35 (indicated by vertical line) and either 500 sites (panel A) or 5000 sites (panel B). Panels C and D show the posterior distributions when the actual rate is either much higher, $r = 2.17$ (panel C) or much lower, $r = 0.09$ (panel D).

different than the topology used to generate the data. The posterior mean of α in the random trees is smaller in all the random trees. As α is inversely proportional to the degree of rate variation, the random trees have a higher heterogeneity of rates among sites. Plotting the posterior mean of rates of the first 150 sites for a 75 taxa and 1000 sites for a dataset generated with a GTR model and $\alpha = 0.5$, I observed a tendency to underestimate the lowest rates and to overestimate the highest rates (see Fig. 2.7). Looking at the posterior distributions of rates, I have noticed that the mode is clearly shifted for some sites. A few examples of such sites are shown in Fig. 2.8, where the distribution at a site using the correct tree (line 1) is contrasted with the distribution obtained with 9 random trees (lines 2-10).

	25taxa1000sites		50taxa1000sites		75taxa1000sites	
	α	R	α	R	α	R
tree 1	**0.24**±0.01	**0.87**	**0.20**±0.01	**0.95**	**0.19**±0.01	**0.92**
tree 2	0.22±0.01	0.81	0.16±0.01	0.84	0.16±0.01	0.87
tree 3	0.21±0.01	0.77	0.16±0.01	0.86	0.15±0.01	0.83
tree 4	0.21±0.01	0.81	0.16±0.01	0.87	0.15±0.01	0.86
tree 5	0.21±0.01	0.83	0.16±0.01	0.82	0.15±0.01	0.84
tree 6	0.21±0.01	0.82	0.17±0.01	0.84	0.16±0.01	0.83
tree 7	0.21±0.01	0.79	0.16±0.01	0.81	0.15±0.01	0.84
tree 8	0.21±0.01	0.81	0.16±0.01	0.82	0.15±0.01	0.86
tree 9	0.22±0.01	0.80	0.16±0.01	0.86	0.16±0.01	0.82
tree 10	0.20±0.01	0.78	0.17±0.01	0.84	0.16±0.01	0.83
tree 1	**0.54**±0.04	**0.85**	**0.51**±0.03	**0.92**	**0.51**±0.03	**0.93**
tree 2	0.35±0.02	0.80	0.33±0.01	0.82	0.33±0.02	0.81
tree 3	0.35±0.02	0.76	0.33±0.01	0.82	0.32±0.01	0.79
tree 4	0.36±0.02	0.74	0.33±0.01	0.82	0.34±0.01	0.84
tree 5	0.37±0.02	0.80	0.35±0.01	0.83	0.33±0.01	0.82
tree 6	0.35±0.02	0.77	0.34±0.02	0.82	0.33±0.01	0.80
tree 7	0.39±0.02	0.79	0.34±0.01	0.82	0.33±0.01	0.82
tree 8	0.35±0.02	0.79	0.33±0.01	0.82	0.35±0.01	0.83
tree 9	0.39±0.02	0.80	0.34±0.02	0.81	0.34±0.01	0.82
tree 10	0.40±0.02	0.81	0.33±0.01	0.82	0.34±0.01	0.80
tree 1	**0.94**±0.08	**0.77**	**0.99**±0.07	**0.84**	**1.04**±0.06	**0.88**
tree 2	0.46±0.03	0.66	0.50±0.02	0.69	0.50±0.02	0.67
tree 3	0.45±0.02	0.65	0.51±0.02	0.68	0.52±0.02	0.70
tree 4	0.47±0.03	0.67	0.47±0.02	0.65	0.53±0.02	0.70
tree 5	0.46±0.03	0.69	0.53±0.03	0.68	0.50±0.02	0.67
tree 6	0.43±0.02	0.65	0.50±0.02	0.68	0.54±0.03	0.70
tree 7	0.45±0.02	0.66	0.47±0.02	0.64	0.51±0.02	0.69
tree 8	0.48±0.03	0.68	0.49±0.02	0.67	0.53±0.02	0.69
tree 9	0.44±0.03	0.64	0.49±0.02	0.67	0.56±0.03	0.71
tree 10	0.46±0.02	0.67	0.49±0.02	0.66	0.53±0.02	0.71

Table 2.2: Estimates of α and the correlation coefficient R calculated between estimated and true site specific rates with correct and incorrect trees. The horizontal line separates the sets generated with different α. The results obtained using the correct tree (T1, in bold) and 9 random trees (T2-T10) are shown.

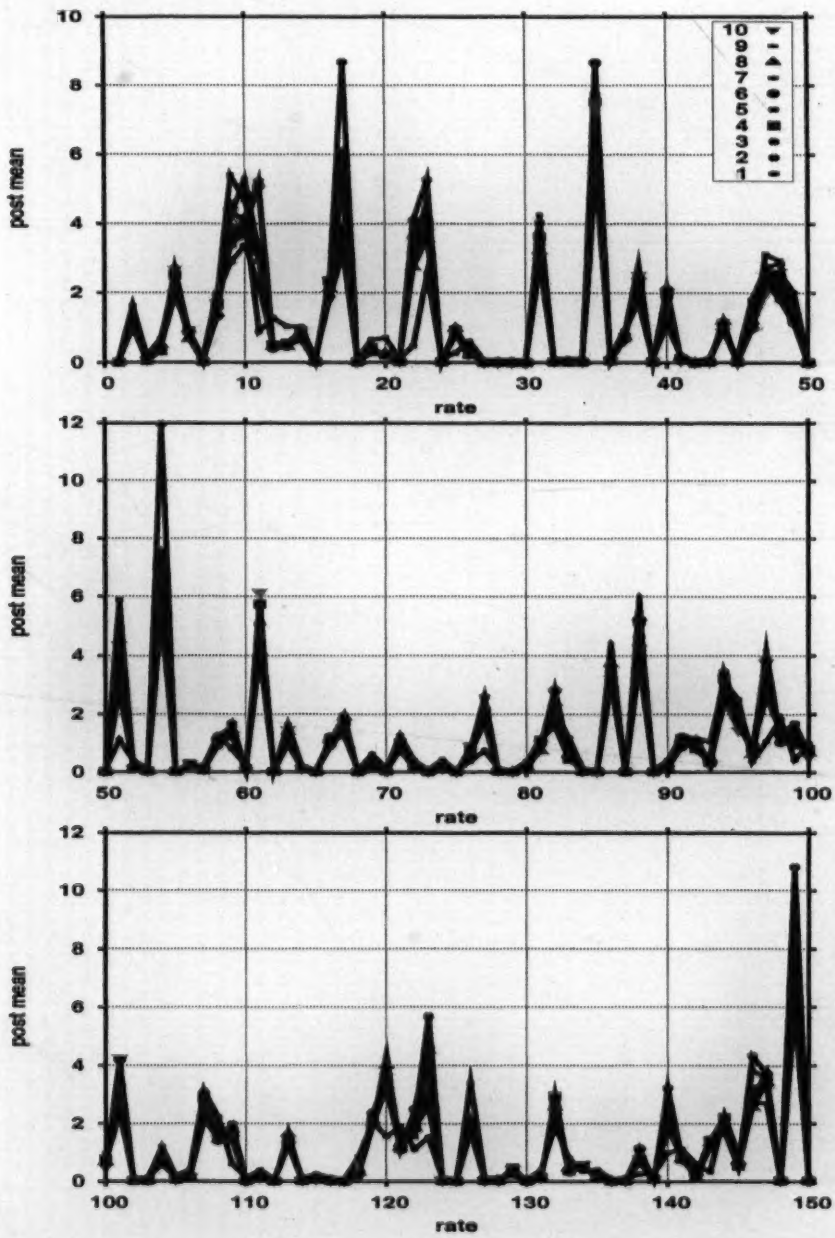


Figure 2.7: Posterior mean of first 150 sites using the correct tree (tree 1, black line) and 9 random trees (label 2-10 in legend) for the dataset of 75 sequences and 1000 sites generated with $\alpha = 0.5$ and the GTR DNA substitution model.

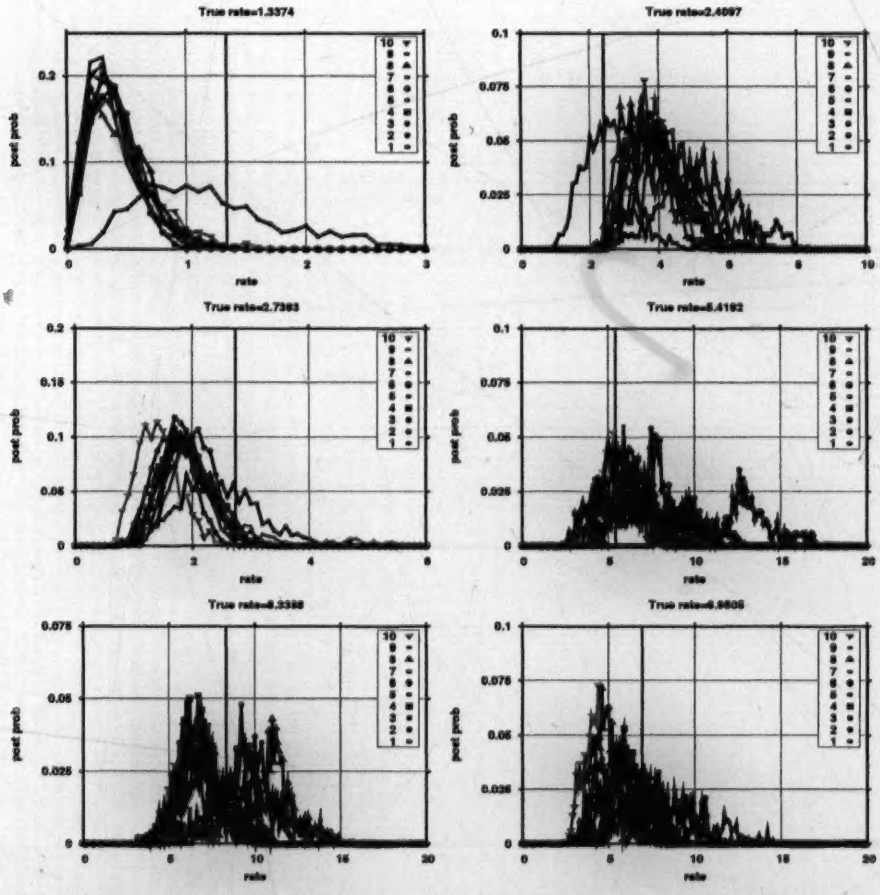


Figure 2.8: Posterior distribution of substitution rate of a few selected sites at which the random topologies gave different estimates. Black vertical line corresponds to the distribution of substitution rate at that site obtained using the correct topology while distributions marked from 2 to 10 are for the substitution rate at site obtained using an incorrect topology. Dataset of 75 taxa and 1000 sites were generated with $\alpha = 0.5$ and analyzed using a GTR model.

2.3.4 Testing extreme distributions of substitution rates

Very high rate variation

The gamma distribution takes the shape of an exponential distribution when the α parameter is less than 1, changing completely the biological representation of substitution rates in comparison with a gamma distribution with a supraunitary α. As α decreases, the rates are increasingly clustered near 0 and an α less than 0.1 means that at least half of the sites have substitution rates less than 0.1 and that less than 5-10% of the sites have extremely high rates. Datasets having such properties show a high degree of conservation for the majority of the sites with few exceptional candidate sites for positive selection. I am interested to investigate the behavior of the method I proposed in such extreme situations. The discretization of the gamma distribution, implemented in PAML, is also analyzed.

A dataset of 1000 sites and a random topology was generated in BYPASSR-gen using the GTR substitution model. The parameters of the GTR model for these datasets are $a = 1, b = 2, c = 3, d = 0.1$ and $e = 0.2$. The nucleotide frequencies are $\pi_T = 0.1$, $\pi_C = 0.2$, $\pi_A = 0.3$ and $\pi_G = 0.4$. The parameter for the prior distribution on rates is set to $\alpha = 0.1$. The λ parameter is set to 100, meaning an average branch length of 0.01. A small α has biological meaning only for short trees, as most of the sites do not change at all. For example, among the 1000 sites with rates drawn from the gamma distribution with $\alpha = 0.1$ and for a tree with 25 taxa and length 0.49, there are 130 site patterns. A site pattern is obtained by combining sites that show the same nucleotides in all species, therefore in my example, the data matrix has 130 different columns of configurations. The same set of 1000 rates was used to generate sequences for a random topology of 50, 75 and 100 taxa.

The simulation results in the previous sections show a very strong concordance between the rate estimates of BYPASSR and BASEML with 50 categories and, obviously, similar estimates of α. With the present choice of parameters, the results are somewhat surprising. In all the analyzed datasets, BYPASSR overestimates α. The posterior mean of α can be more than doubled when compared with the true value. The data is not very informative when the rates are low and the prior has a larger effect on point estimates resulting in an upper bias, however the coverage of the credibility interval is correct.

One advantage of a Bayesian method is the availability of the posterior distribution for the parameters of interest. The mean, mode and 95% highest posterior density interval (95%HPD) are common statistics to evaluate these distributions. To calculate the 95%HPD interval, the samples collected from the chain (after convergence) are binned. The bins are ranked according to their probabilities (from high to low) and the probabilities are summed. When the cummulative probability reaches 0.95, the bins corresponding to the lowest and the highest parameter values are taken as the limits of the credible interval. For two randomly chosen sites the 95%HPD intervals were constructed. The parameter values corresponding to the bins below the black line in Figure 2.9 are not included in the HPD credible set. The ideal is a narrow 95%HPD interval around the mean of the distribution. When the interval is spread across many bins, the variance of

the estimates is large.

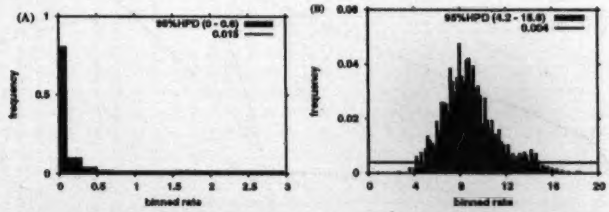


Figure 2.9: Posterior distribution of rate at two random sites binned to calculate 95% highest posterior density. The bins with frequency less than 0.015 (panel A) and 0.004 (panel B) are not included in the HPD credible set.

I calculated the HPD for all the site rates. Figure 2.10 shows the first 100 site rates with the corresponding 95%HPD estimated from the data of 25 and 100 sequences. The lengths of the 95% HPD, plotted as error bars around the mean of the distributions, are greatly reduced when the inference was done on a dataset with many sequences. However, in all the datasets I analyzed the majority of the true rates and the BASEML marginal rates are in the 95%HPD interval. For example, in the two datasets in Fig. 2.10, 96.7% and 98.4% of the true rates are in the confidence interval for the site rates. The performance of rate estimates is much better with the dataset of 100 sequences (Fig. 2.11 panels A) and outperforms BASEML (Fig. 2.11 panels B).

Analyzing the performance of BASEML, I noticed that 50 categories are no longer enough to fit such extreme rates. This is indicated by the presence of cut-off lines above which the rates are not allowed. The discretization of the gamma distribution is done by splitting it into 50 probabilistically equal categories. The boundaries for each category, as output by BASEML, are shown in Fig. 2.12. While 17 categories are reserved for rates $< 10^{-5}$, 26 categories for rates between 10^{-5} and 1.20, 5 for rates between 1.20 and 5.87, all rates above 5.87 have to be allocated one of the two remaining categories. It is clear that the last two categories are mostly affected by the discretization approximation. Increasing the number of sequences, the boundaries between categories are slightly modified and there is still a single category for rates above 10 and one category for the rates between 5 and 10. This observation confirms my previous findings that the small and intermediate rates are overrepresented to the detriment of the highest rates in a dataset. As α gets smaller, more categories are generated for rates approximately near 0 ignoring the highest rates. This improves estimation of the smallest rates, but produces worse estimates for the few extremely high rates.

BYPASSR seems to overestimate the smallest rates, α is overestimated and the rate heterogeneity is slightly reduced (Table 2.3). Fig. 2.13 shows the posterior mean of branch lengths, the maximum likelihood estimates, and the true values.

dataset seqs.	BYPASSR α	BASEML α	BYPASSR TL	BASEML TL	TRUE TL
25	0.365±0.023	0.089	0.375±0.03	0.426	0.493
50	0.286±0.015	0.106	0.950±0.05	1.051	1.153
75	0.265±0.013	0.099	1.076±0.06	1.149	1.251
100	0.240±0.011	0.101	1.692±0.13	1.951	2.030

Table 2.3: Estimates of α and tree length (TL) for datasets generated with $\alpha = 0.1$.

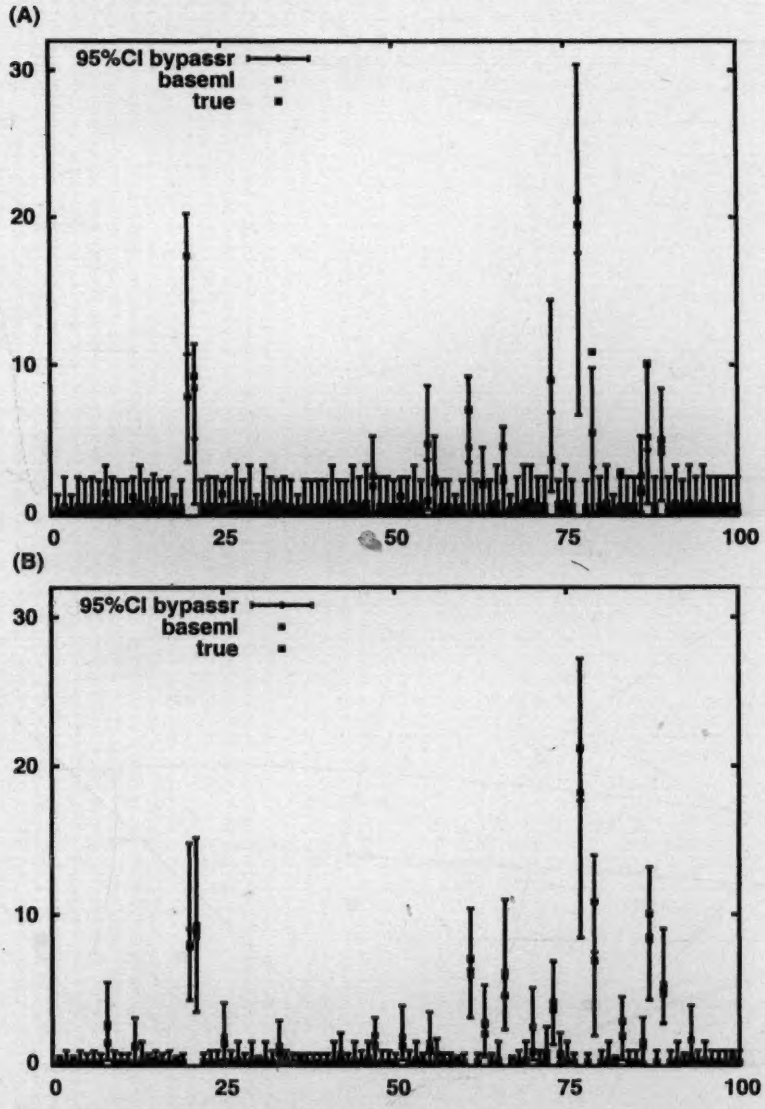


Figure 2.10: 95% Confidence intervals for posterior mean of substitution rates for the first 100 sites in the dataset of 25 (panel A) and 100 sequences (panel B).

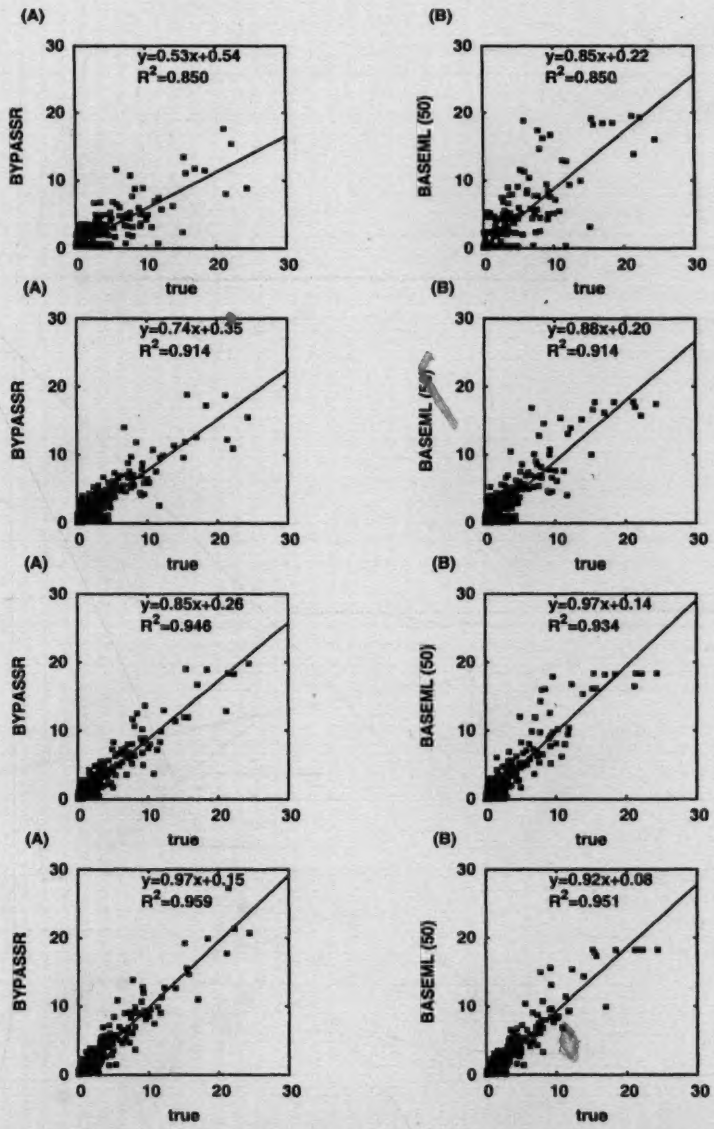


Figure 2.11: Correlation coefficient and regression equations calculated for the true rates and the posterior mean of rates of BYPASSR (A panels) and marginal rates of BASEML (B panels). Datasets of 25, 50, 75 and 100 sequences (from top to bottom) and 1000 sites having substitution rates chosen from a continuous gamma distribution with $\alpha = 0.1$.

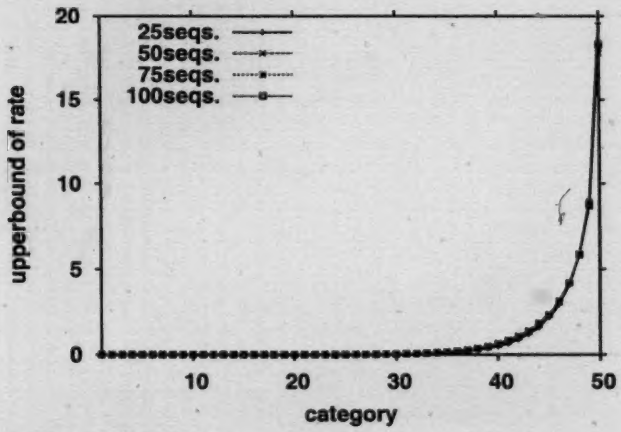


Figure 2.12: Boundaries for the 50 categories used by BASEML to approximate the gamma distribution with $\alpha = 0.1$.

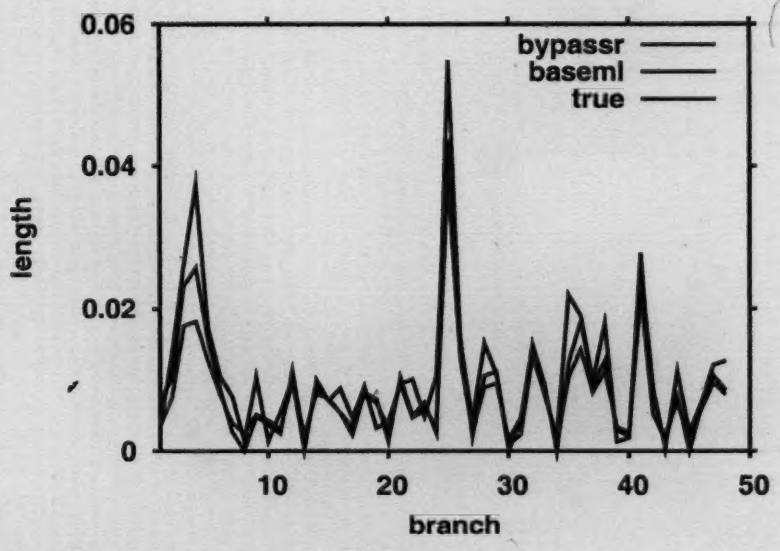


Figure 2.13: BYPASSR and BASEML (50 cat.) branch lengths estimates for 25 sequences. True branch lengths are shown in black.

Low rate variation

Rate constancy of substitution rates among sites is represented by a α parameter greater than roughly 2. There is no distinction between the cases when all the rates are very small or all the rates are very high, both cases having the same α. The gamma distribution is parameterized such that the mean is 1 by setting the scale parameter β to the inverse of α. A value of α anywhere in the interval from 2 to 12 gives a highly concentrated distribution around the value 1. Such a dataset suggests that the DNA region is selectively neutral. I performed a simulation study to assess whether the method is able to identify such situations.

The parameters of the DNA substitution models are the same as in the previous simulations. The same set of site rates simulated from a continuous gamma distribution with $\alpha = 2.5$ is used for generating arrays of sequences with 25, 50, 75 and 100 taxa. The branch length prior with $\lambda = 20$ assures a tree long enough to accommodate many high rates. The gamma distribution is parameterized such that the mean is 1. A typical site from this dataset has the posterior distribution of substitution rate as illustrated in Fig. 2.15.

The estimates of rates obtained with BYPASSR and BASEML are again contrasted with their true values and their correlation coefficient and regression equation are shown in Figure 2.14. As the majority of the sites have reduced amount of information caused by uniformly high rates, increasing the number of sequences at 100 has less impact than for a smaller α. In all the datasets, with BYPASSR and BASEML, as well, α was successfully found to be close to 2.5.

2.3.5 Analyzing independence among sites

Independence of substitution process among sites is a common assumption in phylogenetic inference and I incorporate this assumption in my model as well. The underlying hypothesis, in the biological context, is that the evolution at a site is not influencing the evolution of any other site. Therefore, the substitution rates should not be correlated among sites. This assumption can be evaluated by estimating the variance-covariance matrix of the substitution rates among sites. Independent samples S of substitution rate r_i at each site i from the converged chain are collected at a given interval. The mean and the variance of substitution rate at each site are calculated $\overline{r_i} = \sum_{k=1}^{S} r_{i_k}/S$ and $\text{var}(r_i) = \sum_{k=1}^{S}(r_{i_k} - \overline{r_i})^2$. The covariance between site i and j is

$$\text{cov}(r_i, r_j) = \sum_{k=1}^{S}(r_{i_k} - \overline{r_i})(r_{j_k} - \overline{r_j})/S.$$

The covariance matrix for l sites is written as

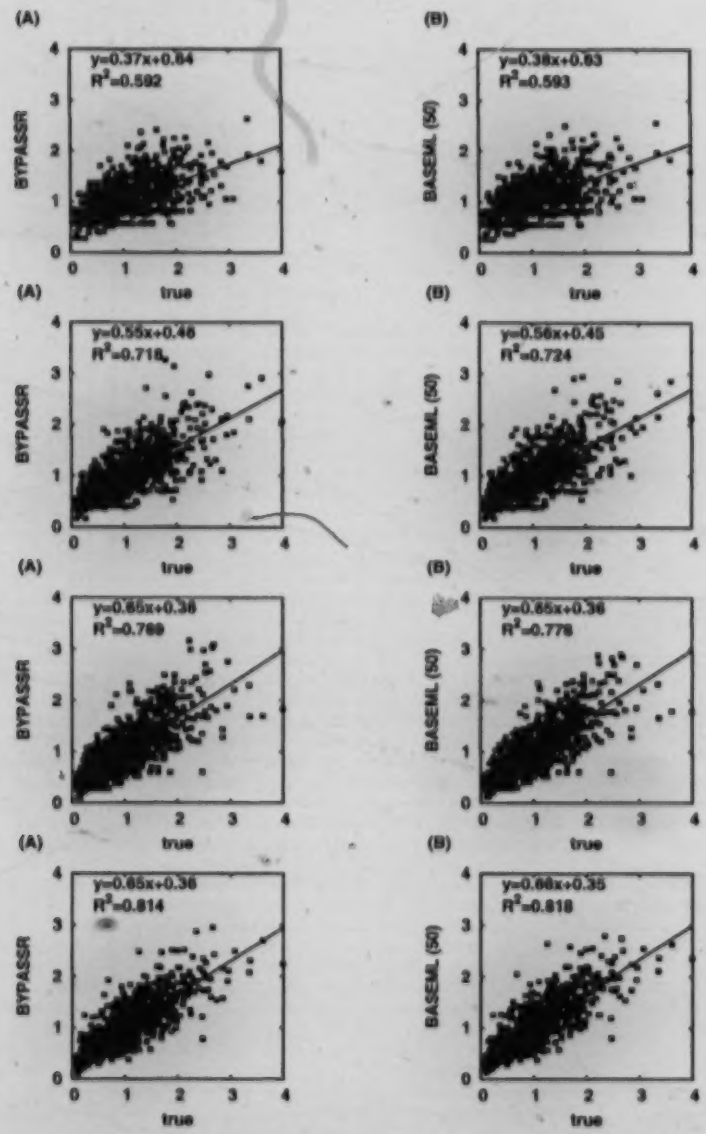


Figure 2.14: Correlation coefficient and regression equations calculated for the true rates and the posterior mean of rates of BYPASSR (A panels) and marginal rates of BASEML (B panels). Datasets of 25, 50, 75 and 100 sequences (from top to bottom) and 1000 sites having substitution rates chosen from a continuous gamma distribution with $\alpha = 2.5$.

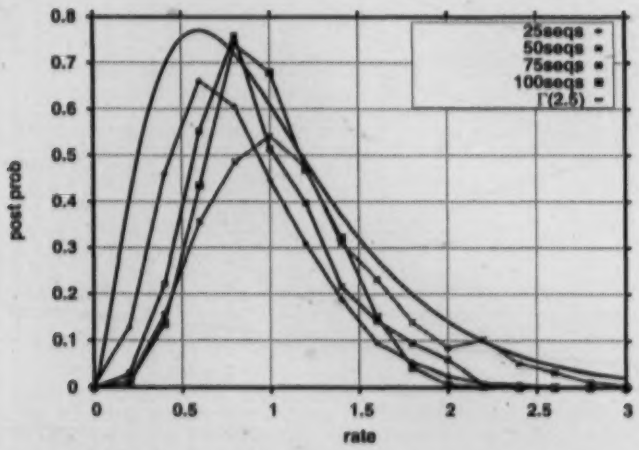


Figure 2.15: Posterior distribution of the rate for a typical site when $\alpha = 2.5$ when the number of sequences vary from 25 to 100. The black line shows the prior distribution, Gamma with parameter $\alpha = 2.5$.

$$\mathrm{var}(\mathbf{r},\mathbf{r}) = \begin{pmatrix} \mathrm{cov}(1,1) & \mathrm{var}(1,2) & \ldots & \mathrm{cov}(1,l) \\ \mathrm{cov}(2,1) & \mathrm{cov}(2,2) & \ldots & \mathrm{cov}(2,l) \\ \cdot & \cdot & & \\ \cdot & & \cdot & \\ \cdot & & & \cdot \\ \mathrm{cov}(l,1) & \ldots & & \mathrm{cov}(l,l) \end{pmatrix}$$

The covariance between two parameters is influenced by the magnitude of the standard deviation of the two parameters. To get a better indication of the covariation of the two parameters, the covariance is scaled to obtain the correlation coefficient $\rho(r_i, r_j) = cov(r_i, r_j)/\sqrt{\mathrm{var}(r_i)\mathrm{var}(r_j)}$. A correlation coefficient equal to 1 means that the parameters for which it was calculated are perfectly correlated, while $\rho < 0$ indicates negative correlation. In this way, the correlation coefficient is calculated for all the elements of the covariance matrix and the correlation matrix is obtained.

Datasets of 25, 50 and 75 taxa and 2000 sites were generated with $\lambda = 35$, $\alpha = 0.5$, relative rates of the GTR model (0.5,1,1.5,2.5,5) and nucleotide frequencies ($\pi_T = 0.1, \pi_C = 0.1, \pi_A = 0.1, \pi_G = 0.4$). The means, variances and covariances were calculated for 2000 samples. The correlation matrix for the 37 sites with mean substitution rate > 4 is shown in the plot Fig. 2.16 (panel A). The site independency is also evident for the slowly evolving sites with rates < 0.1 (panel B). Each tile is the correlation coefficient between the rates of the site on x axis and y axis. The correlation coefficient between any two sites should be greater than 0.5 to be considered meaningful. No structure in the matrix is observed in any of the analyzed simulated datasets.

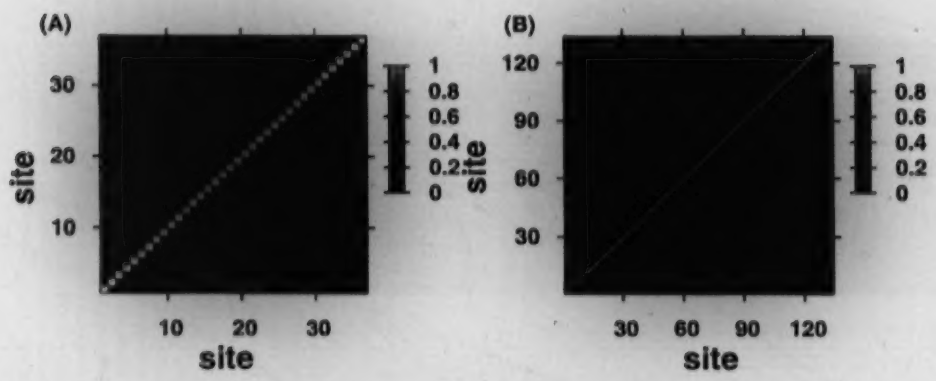


Figure 2.16: Correlation matrix for the 37 sites with mean substitution rate > 4 (panel A) and for the 134 sites with mean rate < 0.1 (panel B).

2.3.6 Influence of branch length prior on parameter estimates

To determine the influence of the branch length prior on the estimated branch lengths under my model, I compared the mean posterior branch lengths using uniform and exponential prior. Datasets of 25, 50 and 75 taxa are generated with $\lambda = 35$ and $\alpha = 0.5$. The same initial values and number of iterations are used for the analysis of the two cases. The summary of the runs is presented in Table 2.4. The use of uniform prior tends to produce longer branch lengths with a small decrease of the posterior means of substitution rates.

The benefits of using an exponential prior are evident when short trees are analyzed. The common parameterization of the substitution models gives the branch lengths in expected number of substitution on time unit. In general, the real datasets have subunitary branch lengths as less than 1 substitution is expected per unit time. The exponential prior proposes values closer to biological reality and it is more computationally efficient. I, therefore, decided to use the exponential prior in all later analyzes.

2.3.7 Implementation of the pruning algorithm

I implemented the pruning algorithm to test the computational efficiency in comparison with my approach using data augmentation. In my approach, nucleotides are explicitly assigned at the internal nodes and their transition probabilities calculated and stored when parameters of the instantaneous rate matrix are changed. To implement the pruning algorithm, I eliminate the step in which nucleotides are assigned and instead I integrate over all 4 nucleotide states. Data is augmented only through the presence of explicit substitution events along the branches of a phylogenetic tree. The pruning

dataset	λ true	λ est.	TL true	TL est
25taxa2000sites	35	37.132 ±5.59	1.34	1.32±0.05
50taxa2000sites	35	33.150 ±3.55	2.96	2.99±0.11
75taxa2000sites	35	32.273 ±2.90	4.64	4.60±0.15
25taxa2000sites	NA		1.34	1.42±0.05
50taxa2000sites	NA		2.96	3.42±0.13
75taxa2000sites	NA		4.64	5.61±0.28

Table 2.4: Posterior mean of λ and tree length using exponential (upper part) and uniform prior for branch lengths (lower part).

algorithm requires, at each node and site of a tree, that a vector of 4 likelihoods calculated using the conditional likelihoods at immediate descendant nodes be stored. If I consider the simple Jukes Cantor DNA substitution model with the instantaneous rate matrix calculated only once at the beginning of the run, the conditional likelihoods at a node change only when the number of substitutions on the branch of one immediate descendant node is modified. If the branch happens to be at the tip, all the conditional likelihoods from that node to the root must be recalculated. Therefore, increasing the number of taxa creates deeper trees, longer paths to the root and more nodes for which conditional likelihoods have to be calculated every time the number of substitutions changes. If a more complex DNA substitution model is used instead, a change of the instantaneous rate matrix causes the conditional likelihoods for all the nodes and sites to be recalculated.

I ran BYPASSR with and without the pruning algorithm implemented for several datasets. A run time of less than 30 minutes on a Dual Core AMD Opteron processor was enough to achieve convergence of BYPASSR using data augmentation for nucleotides at the internal nodes, while convergence of the pruning algorithm implementation required more than 2 hours (see Figure 2.17). For the plots in Fig 2.18 convergence was not reached even after impressively long run for a such small dataset. Increasing the number of sites or nodes increases dramatically the computation time because at each iteration the number of substitutions is changed at a branch or a site, requiring a cascade of calculations up to the root node. Although my implementation of the pruning algorithm may not be fully optimized for efficient computation, its longer running time prompted me to use data augmentation exclusively in the analysis.

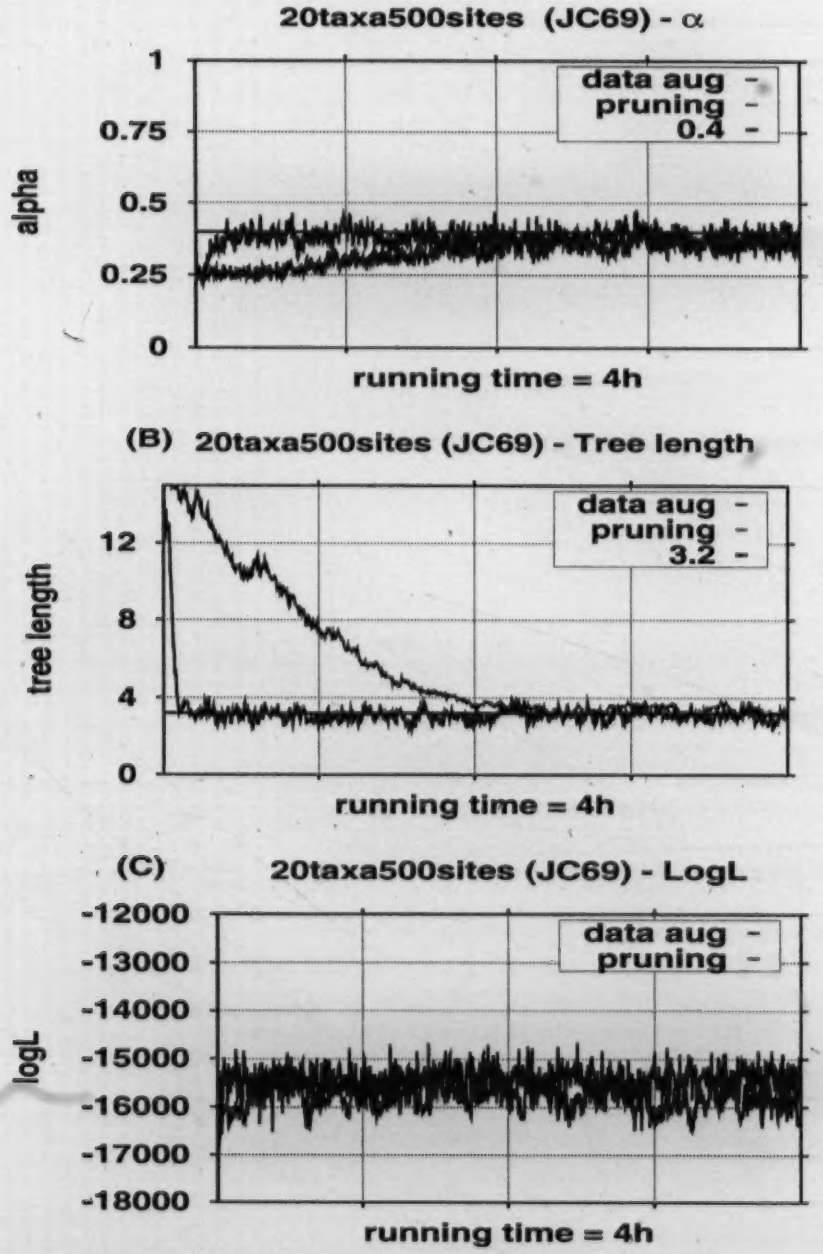


Figure 2.17: Trace plots of the α, the tree length and the marginal tree log likelihood, for the dataset of 20 taxa and 500 sites, as a function of run time with equal run time using either a pruning algorithm or data augmentation with the uniformization technique.

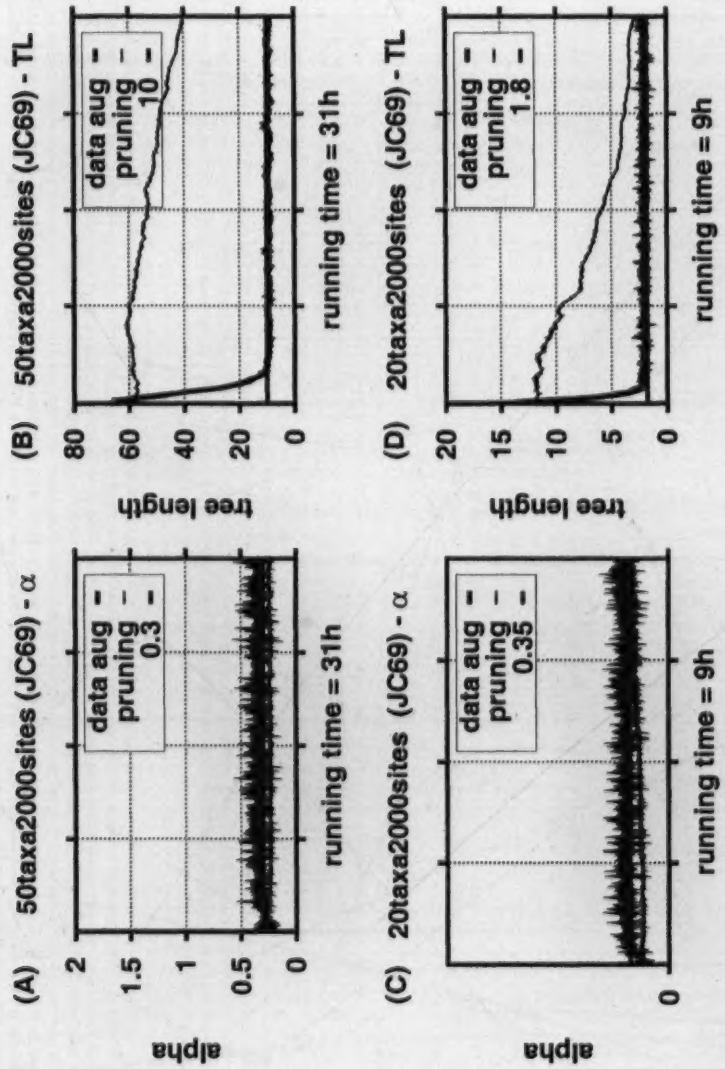


Figure 2.18: Trace plots of α and tree length for two datasets with either 20 or 50 sequences and 2000 sites. Equal run time is used for both datasets analyzed under JC69 substitution model with either pruning algorithm or data augmentation with the uniformization technique.

2.3.8 Assessing convergence

Monitoring and detecting convergence in a Markov chain is a difficult task. The simplest method to assess convergence is using time series plots. Samples are collected during each burn-in run. The chain is assumed to be converged if a perfect horizontal band is formed along the stabilized value. Another approach is to perform multiple independent runs each initialized with different random values. Multiple chains converging to the same distribution is strong indication that the target distribution was reached. Following this idea, a parallel version of the program that allows multiple simultaneous runs was written.

Because convergence is very important for obtaining reliable estimates, BYPASSR treats differently the initial part of the run in which the parameters are not stabilized (burn-in) and the chain after convergence when samples are collected for the parameters of interest (sampling). The user can perform an initial run and the program stops to verify convergence. Samples collected during the burn-in phase are discarded. At this time the convergence can be verified by plotting the saved samples of α, tree length, tree likelihood and a convergence parameter, an ad-hoc procedure to assess convergence. The sites that share the same pattern (i.e. sites with identical nucleotides across lineages) are stored and the most common pattern is chosen. Substitution rates for all sites with the same pattern are binned into 200 bins each of width 0.1. A maximum of 20 random sites are selected from the most common pattern and their site-specific rate distributions are compared pairwise by taking the difference of each pair of bins. For example, two sites, x and y may have the distribution of rates as shown in Fig 2.19. Because sites with the same pattern are expected to have the same posterior distribution, the value of R_{pair} should be close to 1 if convergence is achieved and the two distributions fully overlap. If $R_{pair} = 0$, the distributions have not a single bin in common.

$$R_{pair} = 1 - \frac{1}{2} \sum_{i=1}^{\text{total bins}} |x_i - y_i|$$

The same formula is applied to all 10 pairs of distributions for the chosen 20 sites and the total sum R is printed. A value of R stabilized somewhere close to 1 is a strong indication that the chain has converged. An example of how R evolves over time is shown in Fig. 2.20. The samples collected during burn-in for tree length and the log likelihood of the tree are presented in Fig. 2.21 and 2.22, respectively.

Interestingly, in all the simulations I ran and all the datasets I analyzed, marginal likelihood reached stationarity much earlier than the other parameters in the chain and does not seem to be a reliable indicator of convergence. However, because the rate of convergence of the MCMC method and the number of samples from the chain that are needed for accurate inferences will vary depending on the specific data, initial parameter values used for the chain, etc, it is difficult to compare the computational efficiency with that of an exact likelihood calculation without carrying out extensive simulation studies.

BYPASSR has the ability to infer the substitution rates by using a continuous Gamma

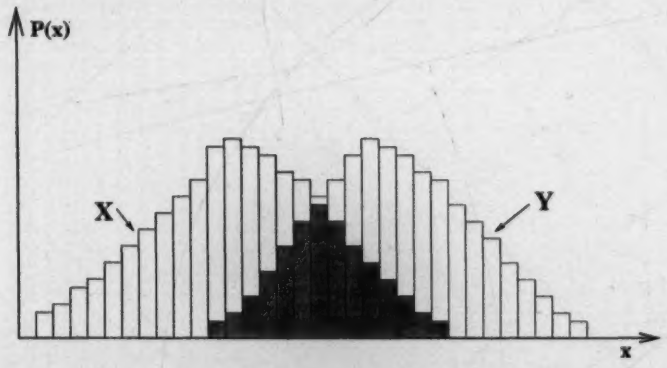


Figure 2.19: Calculating the degree of overlapping of substitution rate distribution at two sites. The area inside the gray zone gives the value of R.

distribution to account for their variation across sites. The inference of substitution rates is done using the Bayesian Markov chain Monte Carlo method. Together with the substitution rates, branch lengths, parameters of the DNA substution model, nucleotides at the internal nodes of the tree are also inferred. The statistical perfomance of the program was assessed by analyzing various data sets that are commonly observed (s.a. tree length in a certain range) or extreme situations (s.a. subtitution rates highly variable across sites). By comparing the results obtained by BYPASSR with the estimates of another implementation constructed on a similar parameteric ground, I validated the method. The similaries between the two methods are obtained for all the parameters except the substitution rates, for which BYPASSR did significantly better. This gave me confidence to proceed to the analysis of empirical datasets. The next chapter is dedicated to this analysis.

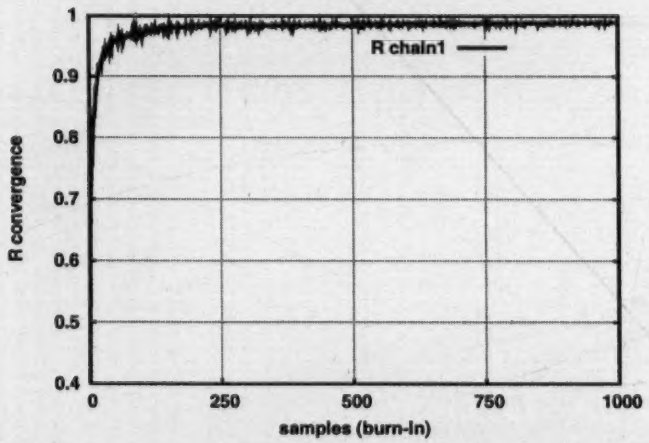


Figure 2.20: R during burn-in for a dataset of 45 sequences and 750 sites.

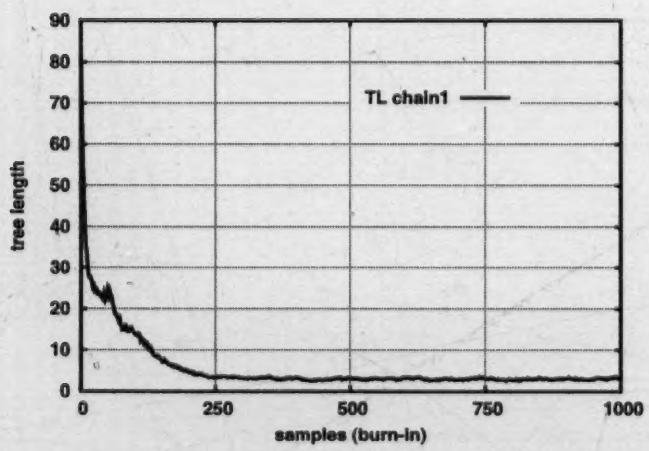


Figure 2.21: Tree length during burn-in for a dataset of 45 sequences and 750 sites.

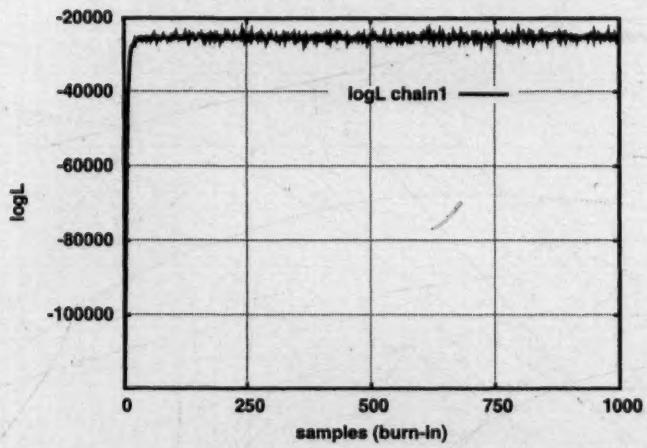


Figure 2.22: Tree log Likelihood during burn-in for a dataset of 45 sequences and 750 sites.

Chapter 3

Site-specific rate variation in several genes

The accurate inference of rates obtained from simulated data sets provides confidence in the use of BYPASSR for the analysis of published datasets. Some of the datasets required new features to be added to BYPASSR, or additional codes to be written. The use of the fasta format for DNA or protein sequences and the newick [101] representation of trees has become standard in phylogenetics. BYPASSR deals with both and allows data with gaps to be input. Perl and bioperl scripting (web reference [102]) allowed me to perform customized searches for patterns, or to download multiple sequences in an automated way. The graphical representation of tens or hundreds of pages of numbers is essential in any statistical analysis involving thousands of variables. GNUPLOT is a very flexible command-line software allowing nice visualization of the data [103]. I included the feature that GNUPLOT scripts are automatically generated at the end of each run. All the statistical analysis of the data is included in BYPASSR and therefore does not require the use of additional software.

I have chosen six datasets to illustrate the important findings of my method (Table 3.1). Analyses of the performance of the method have shown that longer sequences do not improve estimates of site-specific rates, but larger numbers of sequences generate more accurate posterior means. The length and number of sequences in each analyzed dataset varied greatly. The Japanese Encephalitic Virus has only 20 sequences, but more than 10000 nucleotides per sequence, while the mammalian cytochrome *b* has an impressive 688 species. Parametric methods for phylogenetic inference with complex substitution models, in particular the maximum likelihood method, can cope very well with large numbers of sites. The number of sites used in such methods of analysis is reduced to the number of unique patterns. My Bayesian MCMC approach does not allow site patterns to be treated in this way because of the implementation of the continuous gamma distribution and the sampling algorithm. The sites with identical pattern have the same distribution of their substitution rates, but they cannot take equal values simultaneously in the chain. On the other hand, no simplification can be done for parameters related to the branches of the phylogenetic tree, so the maximum likelihood method (ie.

Gene	seqs.	sites used
Japanese encephalitis virus+	20	10298
3-prime noncoding		586
EDN/ECP	18	390
Rabies virus glycoprotein+	35	2069
5-prime noncoding		490
HIV-1 *pol nef*	23	2841
MHC class I	192	810
Cytochrome *b*	688	1147

Table 3.1: Data sets analyzed with BYPASSR.

BASEML) limits the number of input sequences to a few hundred. Theoretically, the method I have proposed has no limit imposed by the number of sequences other than the running time. To my knowledge, a large dataset like the mammalian cytochrome *b* has not been analyzed previously using a continuous distribution of substitution rates at each site.

The most widely used software that can be related to the theory implemented in BYPASSR is MrBayes [104]. The same assumptions hold in both programs: DNA substitution viewed as a Markov process, site independence, no recombination, etc. The site specific rates are modeled in MrBayes by an approximate Gamma distribution as proposed by Yang (1994) [50]. The parameters of the substitution model, branch lengths and site rates are calculated via Bayesian theory instead of using maximum likelihood as in PAML. Codon based models implemented in MrBayes are similar with those in PAML. One major strength of MrBayes is that it accommodates the uncertainty of tree topology. For every tree, a probable combination of parameter values is found. In the end, the tree, with the highest posterior probability is chosen. The set of parameter values in the MrBayes output for a tree with high posterior probability should be close to the BYPASSR estimates, if the topology used in BYPASSR was not accurate. As shown in the theory section, my method can be extended to integrate over tree topology as well. However, the focus of this research is parameter estimation, not tree searching. Although PAML has the ability to do tree searching it does not do it jointly with the integration over all possible combinations of the other parameter values. Results of the maximum likelihood implementation are therefore easier to interpret in validating my results because they are based on a single fixed topology as in BYPASSR analysis.

3.1 HIV-I *pol* gene

The strong correlation of clinical outcome with high levels of genetic polymorphism is one reason why HIV-1 is one of the most sequenced organisms. HIV-1 genes are ideal working material for one interested in selection. Positive selection is known to be an

important factor in the evolution of this virus [105]. The original codon substitution model was tested on a HIV-1 dataset [106]. Nowadays, HIV-1 *gag*, *pol* or *env* genes are often used to validate new methods claiming to detect selection at individual sites. An alignment of 23 isolates (2841 bp) for the HIV *pol* polypeptide (alignment and topology available online [107]) was searched for positively or negatively selected codon sites by three prominent research groups. Yang *et al.* [29] modeled the substitution process among the codons of multiple sequences and estimated the dN/dS ratio and the other parameters of the substitution model with maximum likelihood. They proposed various distributions to describe ω. All the proposed distributions are approximated by discretization allowing a limited number of classes of ω values. With the parameters fixed at the maximum likelihood estimates, using the Bayesian formulation to calculate the conditional probabilities, the codon sites are assign to an ω class according to their posterior probabilities. The PAML program implemented this method and identified 10 sites with posterior probability > 0.95 to have $\omega > 1$ and a few additional with probability > 0.5 (see Table 3.3 ML+NEB). The new distribution of PAML 3.14 contains an improved version of the method that assigns priors to the maximum likelihood estimates, therefore accounting for the uncertainty in the estimates [94] [108]. I have analyzed the HIV-1 *pol* dataset with PAML (CODEML) using the M8 model ($\omega < 1$ from β distribution and one freely estimated category for $\omega > 1$) and different starting ω values as recommended by the author. The codon sites with posterior probability > 0.5 are showed in Table 3.3 (ML+BEB). Another method applied to this dataset uses a maximum likelihood model to find ω and the other parameters of a slightly different codon substitution model than is implemented in PAML, but a likelihood ratio test is used instead of Bayes formula to pinpoint the nonneutral evolving codon sites [109]. The dataset is analyzed under the null hypothesis ($\omega = 1$) and the alternative hypothesis with ω free to vary. A likelihood ratio test applied to each site tests which of the two hypotheses provides better fit. Generating a null distribution (using a Monte Carlo simulation under the null hypothesis) for each site and comparing it with the actual distribution, the P-value at each site can be calculated. The sites that have probability > 0.95 are in the column ML+LRT of the Table 3.3. Huelsenbeck *et al.* used a Bayesian method to analyze this dataset [110]. The topology is not fixed and the parameters of the Yang *et al.* (2000) [29] codon substitution models are estimated using Bayesian conditional probabilities. The essential difference is the treatment of the ω classes. Instead of fixing the number of classes of ω beforehand, the number of classes and the dN/dS ratio for each class are taken from a Dirichlet process prior without assuming a separate distribution for the ω. Interestingly, fewer sites with $\omega > 1$ and probability > 0.95 are found for HIV dataset (Table 3.3 BAYESIAN). The codons with lower probabilities are not available.

I analyzed the same data set with BYPASSR. The chain converged after 10^6 iterations and 2000 samples were collected for summary statistics. Table 3.2 contains the results of the three runs, each initialized with different random values. An exact criterion to identify individual nucleotide sites that undergo selection cannot be formulated because of the difficulty of distinguishing between the mutation and selection processes.

However, several site-specific rate patterns can be defined as criteria to indicate the best candidates for positive selection. A site with high substitution rate at the second codon position, especially if it is higher than the rate at the first and the third codon position, and above the mean of the rates in the dataset, is a good candidate for being under positive selection. A site located at the first position in a codon with the mean substitution rate greater than the posterior mean rate at the third codon position is also a candidate site for positive selection. Sites satisfying these conditions and having more than a specified percent (e.g. 50, 80, 90, 95 or 99%) of the posterior distribution above the rate 1 (mean of rates set by the gamma prior) are potentially positively selected sites.

The list of sites that appear to be under positive selection (and their codons), with the posterior mean and the 95% highest posterior density (or credible interval) are in first column of the Table 3.3. The upper part of the table shows the sites that have $\overline{r_1} < \overline{r_2} > \overline{r_3}$ ($\overline{r_1}$ posterior mean rate at the first codon position, $\overline{r_2}$ at the second codon position and $\overline{r_3}$ at the third) and the probability of $r_2 > 1$ greater than 0.8. In the lower part of the table the sites with $\overline{r_1} > \overline{r_3}$ and the probability of $r_1 > 1$ greater than 0.9 are listed. Probabilities > 0.95 are marked in italics. All previous methods agree that codons 3 and 67 might undergo positive selection. BYPASSR also identifies these codons. Among the 15 nucleotide sites found by BYPASSR, 12 belong to codons in which selection was found by at least one of the maximum likelihood methods. If I also consider the first codon position with the mean rate higher than 1 and also higher than the mean rate at the third codon position, BYPASSR recovers almost all of the codon sites found with high probability by the Bayesian and maximum likelihood methods.

As an extension of Table 3.3, the sites with high mean posterior rates that also satisfy the criteria for being a candidate site are graphicallly illustrated in Fig 3.1. Panel A shows the sites with the highest substitution rate at the second codon position and with a probability of $r_2 > 1$ greater than 0.5. At some sites $\overline{r_2}$ is at least twice as large as the other codon rates, while some codons have all three mean posterior rates well above 1. All methods identified with high probability these sites. BYPASSR finds 3 additional sites 134, 1463, 1508 with high probability. Fig. 3.1 (panel B) shows sites with the highest substitution rates at the first codon position and also being higher than $\overline{r_3}$ with the probability of $r_1 > 1$ greater than 0.9. There are 7 additional sites found with my method 1120, 1183,1195, 2071, 2182, 2581 and 2617.

I have calculated the site-to-site correlation matrix for the substitution rate samples to verify the fit of the site independency assumption of the model. The correlation coefficient for 18 sites (all with $r < 0.6$) was between 0.20 and 0.23, all the other sites having $\rho < 0.2$.

I further investigated how robust the method is when the phylogenetic tree for HIV data set is very likely incorrect, instead of the maximum likelihood tree. I generated 8 random topologies for 23 taxa. I ran the HIV data set with each of these 8 random trees. The expectation is that no significant difference in the posterior distribution of the rates across sites is observed when a random (i.e. incorrect) topology is used. The expectation was met for all the sites with low mean rate,although very few exceptions existed. Two sites found to be positively selected are among them. Figure 3.2 shows

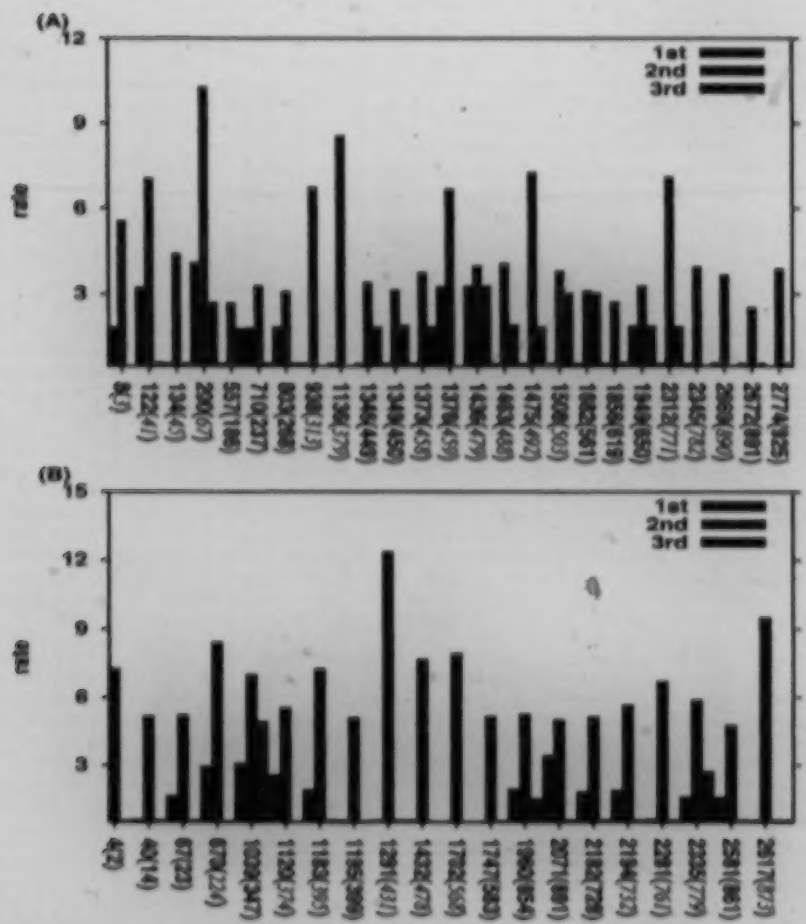


Figure 3.1: Sites (and their codons) with $\overline{r_1} < \overline{r_2} > \overline{r_3}$ and posterior probability P($r_2 > 1$)> 0.5 (panel A). Sites with $\overline{r_1} > r_3$ and > 0.9 P($r_1 > 1$)> 0.9 (panel B). The codons in italic show the posterior probability > 0.95 of either $r_2 > 1$ (top panel) or $r_1 > 1$.

Param.	run I	run II	run III
TL	0.396±0.12	0.396±0.02	0.398 ±0.02
α	0.458±0.02	0.454±0.02	0.455±0.02
λ	113.42±17.79	113.34±18.14	113.66 ±17.99
a	1.187±0.11	1.182±0.11	1.190 ±0.12
b	0.114±0.02	0.112±0.02	0.113 ±0.02
c	0.074±0.01	0.075±0.02	0.073 ±0.03
d	0.262±0.030	0.261±0.02	0.261 ±0.03
e	0.126±0.027	0.127±0.03	0.126 ±0.02

Table 3.2: Posterior mean for model parameters for HIV *pol* sequences.

two typical site rate distributions (site 134 and 938) and two exceptions: site 8 and 2312. At site 8, the rate posterior distributions obtained with the maximum likelihood tree and 8 random trees have two distinct modes, instead of fully overlapping. The rate posterior distributions at site 2312 have a slightly different behavior. Almost every rate posterior distribution gives a different mode than the one obtained using the maximum likelihood tree. As a preliminary conclusion, it seems that for some sites the substitution rate estimate is dependent of the choice of the tree topology used.

BYPASSR		ML+NEB [29]	ML+BEB [94]	ML+LRT [109]	BAYESIAN [110]
2nd pos.(codon)	mean (95%CI)		codon	codon	codon
8 (3)	5.572 (1.4-11.2)	2	2	*2*	*3*
122 (*41*)	7.075 (2-13.6)	*3*	3	*3*	*67*
134 (45)	4.389 (0.6-9.4)	4	4	4	*347*
200 (*67*)	10.29 (3.4-18.2)	*14*	14	*14*	*478*
938 (313)	6.739 (1.8-12.2)	*41*	*41*	*41*	*568*
1136 (*379*)	8.563 (2.6-16.6)	*67*	*67*	67	*779*
1376 (*459*)	6.675 (1.2-13.2)	313	224	313	
1436 (479)	3.986 (0.4-8.6)	*347*	225	347	
1463 (488)	4.066 (0.6-9.2)	*379*	313	*379*	
1475 (492)	7.283 (1.8-13.6)	388	388	*347*	
1508 (503)	3.808 (0.6-8)	431	379	*379*	
2312 (771)	7.119 (1.6-15)	459	388	*459*	
2345 (782)	3.949 (0.4-8.6)	462	*431*	462	
2669 (890)	3.658 (0.6-8.2)	*478*	459	*478*	
2774 (925)	3.877 (0.4-8.2)	506	462	*568*	
		568	*478*	570	
1st pos.(codon)	mean (95%CI)	570	479	*654*	
4 (*2*)	7.279 (1.6-14.2)	583	492	732	
40 (14)	5.181 (1.2-11.2)	654	506	*761*	
67 (23)	5.225 (1-10.4)	*761*	552	*779*	
670 (*224*)	8.438 (2.6-15.6)	*779*	*568*	782	
1039 (347)	7.026 (1.8-14.8)	782	570	890	
1120 (*374*)	5.553 (1.4-12)		583		
1183 (*395*)	7.257 (1.8-14.4)		654		
1195 (399)	5.103 (1.4-10.8)		671		
1291 (*431*)	12.3438 (4.4-21.6)		732		
1432 (*478*)	7.705 (2.2-14.4)		*761*		
1702 (*568*)	7.930 (1.6-15.2)		771		
1747 (583)	5.154 (0.8-10.4)		*779*		
1960 (654)	5.266 (0.8-10.2)		782		
2071 (691)	5.004 (1-10.4)		890		
2182 (728)	5.141 (0.8-11)		892		
2194 (*732*)	5.661 (1-11.8)		894		
2281 (*761*)	6.718 (1.4-13.4)				
2335 (*779*)	5.889 (1-11.8)				
2581 (861)	4.784 (0.8-10)				
2617 (*873*)	9.491 (2.6-18.4)				

Table 3.3: Candidate sites to evolve under positive selection obtained with 4 different methods. The values in italics correspond to the highest probabilities given by the selection criteria of each method. The underlined numbers are the codons found by BYPASSR and at least one of the other methods to be positively selected.

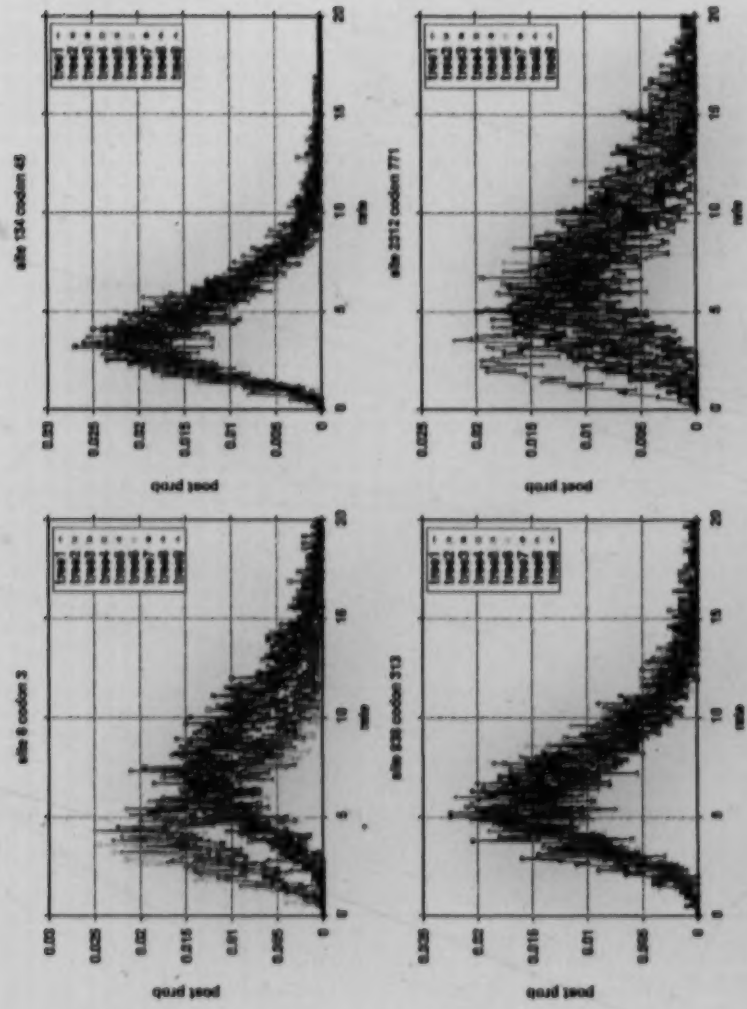


Figure 3.2: Posterior distribution of the substitution rate at site 8, 134, 938 and 2312 when the maximum likelihood tree (line 1 in the legend) and 8 random trees (lines from 2 to 9) are used to analyze HIV data set.

3.2 MHC class I

One of the first studies in which natural selection was demonstrated involved the antigen recognition sites of class I human major histocompatibility genes (MHC) [111]. The interest in understanding the exact location in the gene and the physico-chemical properties of the positively selected sites influenced further development of statistical methods that found such sites in the first place. Instead of assuming one dN/dS rate ratio for all nonsynonymous rates, the later models allow the differentiation of dN/dS according to the amino acid physico-chemical properties [112] [113] [114]. An alignment of 192 sequences and 270 codons (available online at [115]) was analyzed by Yang *et al.*(2002), Sainudiin *et al.*(2002), Yang *et al.*(2005) [108] and Wong *et al.*(2006). All the methods have in common the maximum likelihood framework and they are based on a codon substitution model. The differences among the models are given by how they use the prior knowledge about the positively selected sites, how the data is partitioned, or how the positively selected sites are identified (NEB versus BEB method in the HIV *pol* example).

I analyzed the MHC data set using BYPASSR. The program was run for approximately 20 hours, with 3×10^6 iterations for burn-in and 2×10^6 iterations of sampling during which 2000 sample points were collected for each of the parameters (except the number of transitions and the nucleotides at the internal nodes). As in all of my analyses, the GTR model was used. The results of three runs are summarized in Table 3.4. It is worth noticing the high posterior mean for λ. The upper bound of λ initially set to 200 had to be moved to 400 to fit this dataset. This result suggests a very short tree with an average branch length of 0.003 in the case of the first run. The results are very consistent between runs.

The MHC dataset is one of the cases, as described in the Section 2.3, in which α is very small. The maximum likelihood estimate of α obtained by BASEML with 50 categories is 0.12. The simulation studies have shown that BYPASSR performs comparable with BASEML with 50 categories, but outperforms BASEML when α is very small. A large number of sequences was also found to improve the accuracy of my estimates. The difference in the rate estimates of the two methods are shown in Fig. 3.3. When BASEML uses 5 categories to approximate a Γ distribution with such a small α, all the rates greater than 4.48 are lumped into category 5. Category 4 contains rates between 0.46 and 4.48, while rates smaller than 0.46 are in the remaining 3 categories. Each category has an equal proportion meaning that 60% of the rates are assigned to be < 0.46. Two categories are clearly not enough to fit all the rates above 0.46. The situation is improved when 20 categories are used. The threshold for the rates to be in the 20th category is 11.18. The 7 categories between 0.1 and 4.04, in this case, allow the small and intermediate rates to be better distributed, but there is still one single category for the rates between 4.04 and 11.18. The overestimation of rates in this category is obvious. The increase to 50 categories allows more freedom for high rates to be spread among categories, but there are still problems with rates falling into the two highest categories. There are too many rates in the 49th category (rates between

Param.	run I	run II	run III
TL	1.358±0.09	1.393±0.09	1.412±0.02
α	0.260±0.01	0.259±0.01	0.262±0.02
λ	283.87±25.43	275.45±21.99	273.09 ±26.90
a	1.499±0.17	1.446±0.16	1.454 ±0.16
b	0.968±0.12	0.943±0.11	0.931 ±0.12
c	0.571±0.07	0.559±0.07	0.554 ±0.07
d	0.587±0.08	0.580±0.07	0.587 ±0.07
e	0.741±0.01	0.730±0.01	0.727 ±0.01

Table 3.4: Estimates of shape parameter α of gamma prior on site-specific rates, and 5 relative rates from GTR model, a, b, c, d, e, obtained from the mean of the posterior distribution of three independent runs of the BYPASSR program (run I, II and III) for HIV *pol* sequences.

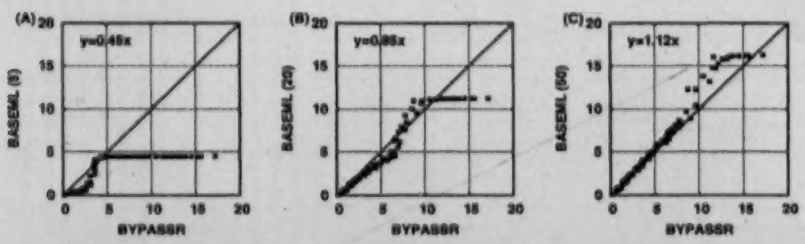


Figure 3.3: Comparison of estimated branch lengths and site-specific rates obtained using BYPASSR and BASEML programs. Panels plot the mean site-specific rates from the posterior distribution generated by BYPASSR (horizontal axis) against estimates of site-specific rates generated using BASEML (vertical axis) with either 5 (panel A), 20 (panel B) and 50 (panel C) rate categories, respectively.

5.72 and 8.27) and too few in the 50th with the cut-off value at 16.24. As shown in my earlier simulations, the rates in the last two categories are most adversely affected by the discretization.

I found the sites with the highest rates at the first and second codon position by calculating how much of the left tail is above 1. When $\overline{r_1} < \overline{r_2} > \overline{r_3}$, I consider such sites to be the potentially evolving under positive selection. Fig. 3.4 (A) shows the sites with more than 80% of the distribution above 1. The corresponding codons in italics show the sites for which the probability is above 0.95. Fig. 3.4 (B) contains the sites located at the first codon position. Among the 808 sites analyzed (sites 31 and 690 were eliminated because of gaps), 19 sites have the rate at the second position as the highest rate in the codon. Sites 455 and 467 have mean rates above 15, the highest rates I have observed in any datasets analyzed. Among the total of 41 codons shown

in Fig. 3.4 with their mean rates, 7 have the relation $\overline{r_1} > \overline{r_3}$ and $\overline{r_2} > \overline{r_3}$. Yang and Swanson (2002) [112] estimated with probability > 0.95 that $\omega > 1$ for codons 9F, 24A, 45M, 62G, 63E, 67V, 69A, 70H, 71S, 77D, 80T, 81L, 82R, 94T, 95V, 97R, 99Y, 113Y, 114H, 116Y, 151H, 152V, 156L, 163T and 167W. Eighteen of these 25 previously found candidates for positive selection were also found by us. The codons that are not in my plot also have high mean rates, but these rates do not exceed the mean rate at the third codon position. This is the case for codon 24 with $\overline{r_1} = 6.65$, $\overline{r_2} = 0.17$ and $\overline{r_3} = 7.14$ or codon 63 with $\overline{r_1} = 14.51$, $\overline{r_2} = 0.15$ and $\overline{r_3} = 15.28$.

The substitution rates are expressed in expected numbers of substitutions per unit time. If I assume that all the substitution rates at the third codon position are synonymous and reflect the rate of mutation, which is assumed not to vary across sites, then I can represent the rates in units of the expected numbers of synonymous substitutions per unit time by setting the rate at all third positions to 1. Everything in the model remains unchanged, except that now I do not propose rate changes at the third codon positions and when I modify α I integrate only over the first and second positions. This different scaling modifies the interpretation of the substitution rates. If $\overline{r_1} > 1$ or $\overline{r_2} > 1$ it means that the site evolves at a faster rate than a synonymous site possibly due to positive selection. A subunitary rate suggests a more slowly evolving site possibly due to negative selection. This interpretation is similar to that of the dN/dS rate ratio.

With this new approach, the posterior mean for tree length is 1.298 and $\alpha = 0.23$. Sites with probability above 0.8 of a rate greater than 1 are shown in Fig 3.5. Sites with probability greater than 0.95 are in italics. All the sites previously found to be positively selected are found by BYPASSR with high confidence. BYPASSR finds 12 additional sites that show a high probability of being candidate sites for positive selection. None of these additional sites were identified by any of the maximum likelihood methods that were used to analyze this dataset. Site 211 belongs to a codon previously identified as a candidate for positive selection, but codons 34, 220, 235, 247, 307, 517 or 580, apparently because they lack a non-zero rate at the second codon position, are not identified as under positive selection by the other methods. The two additional sites at the second codon position found by us, sites 137 and 227, are also missed by the other methods, although they have the same mean rate configuration as sites 281 or 296.

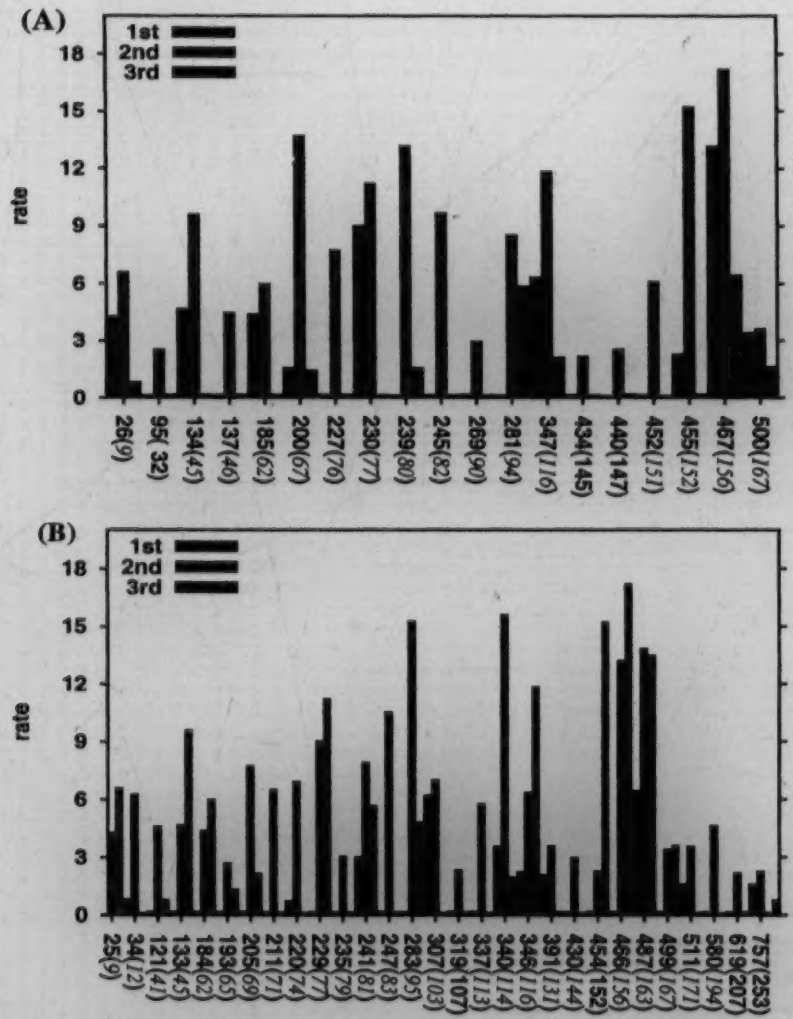


Figure 3.4: Codons in MHC class I gène with $\overline{r_1} < \overline{r_2} > \overline{r_3}$ and posterior probability $P(r_2 > 1) > 0.8$ (panel A). Sites with $\overline{r_1} > \overline{r_3}$ and $P(r_1 > 1) > 0.8$ (panel B). The substitution rates are represented in expected numbers of substitutions per unit time.

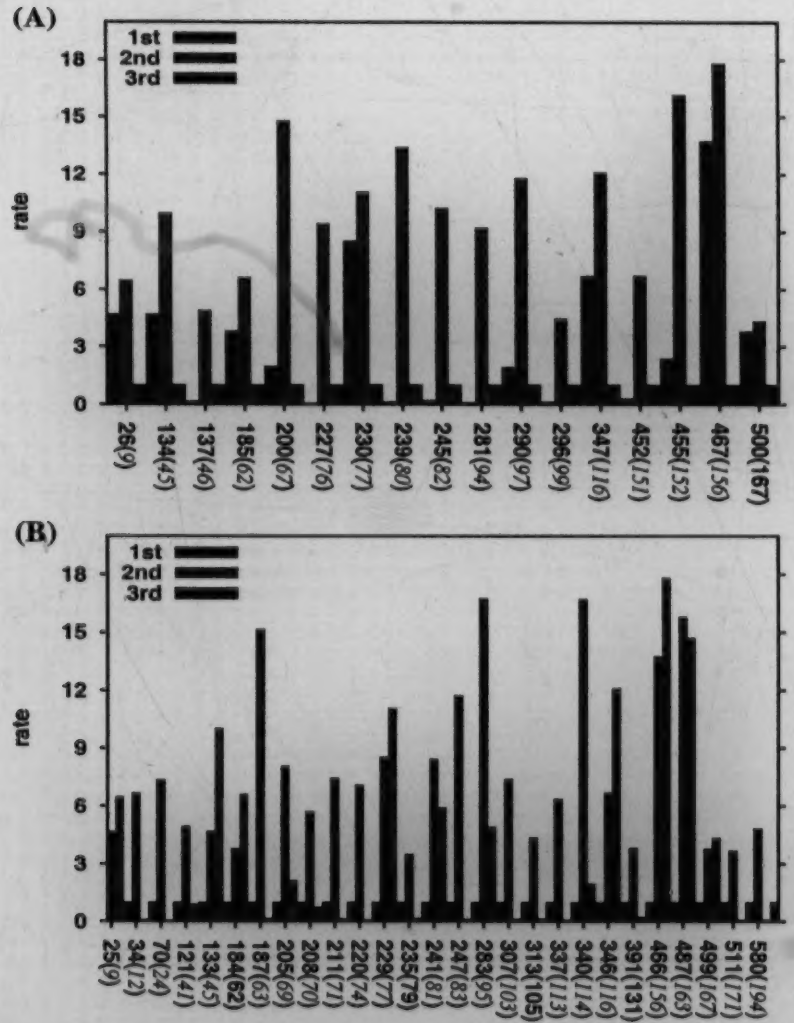


Figure 3.5: Codons in MHC class I gene having posterior probability $P(\overline{r_2} > 1) \geq 0.8$ (panel A). Sites with posterior probability $P(r_1 > 1) \geq 0.8$ (panel B). The substitution rates are represented in expected numbers of synonymous substitutions per unit time.

3.3 Japanese encephalitis virus and 3-prime noncoding

Japanese encephalitis virus (JEV) is a member of *Flaviviridae* with affinity for the central nervous system and is spread by arthropods, usually mosquitoes. Its genome is approximately 11 kb in length with a single open reading frame flanked by approximately 95 noncoding nucleotides at the 5-prime end and 585 noncoding nucleotides at the 3-prime end. Analysis of various JEV isolates has found a region in the 3-prime noncoding zone of variable length (39 to 109 nucleotides) having many polymorphic sites. The rest of the 3-prime noncoding sequence is highly conserved. No clear function has been attributed to 3-prime noncoding region, but some role in viral replication is suspected [116]. The structural genes (capsid C, matrix M and envelope E) are encoded in the first quarter of the genome, while the nonstructural proteins (NS1-NS5) are coded by the remainder [117]. Because of the functions in viral pathogenity (attachment to cellular receptors, membrane fusion etc.), the surface glycoprotein E is the most likely region to have sites that undergo positive selection. The dN/dS rate ratios at each codon site of the E gene were previously estimated with other parametric methods that allow very flexible distributions of ω, but evidence of positive selection was not found [29] [110]. The coding DNA together with the C terminal noncoding segment of 585 nucleotides was analyzed by Wong and Nielsen (2004) using a new method to search for selection in coding and noncoding DNA (see section 1.2.2) [39]. They also concluded that positive selection is not present in the viral genes or in the noncoding region. The alignment of 20 isolates was requested from the authors. I obtained the phylogenetic tree of the coding DNA sequences with the tree searching options in BASEML (PAML) [94] using the GTR model and 5 categories for the gamma distribution of rates (Fig. 3.6).

I am interested in substitution rates in the coding and noncoding regions of the genome. When I tested my method on simulated data, I observed that the effects of discretization performed by BASEML are more visible when the number of sites is increased to 5000. More specifically, it is much harder to distribute, for example 10000 points into a limited number of categories. I compare site-by-site rate obtained with BYPASSR and BASEML with 5, 20 and 50 categories. The three independent chains converged after approximately 5 hours of running time on the university's Opteron cluster. The posterior mean tree length is 0.376 ± 0.009 and $\alpha = 0.530 \pm 0.02$.

The posterior mean of rates (Fig. 3.7) obtained in BYPASSR and the marginal rates from BASEML with 5, 20 and 50 categories for 10,284 sites are binned at the 0.1 interval. A distribution approximated by so many points should have a continuous shape, with the majority of the bins filled. However, because of the small amount of information about site-specific rates contained within only 20 sequences, few bins gathered points. In other words, even if I would have fewer sites with this tree the same bins are formed, but the height of the bars in the histogram would be reduced. As regards BASEML performance, my expectations are met and again, as the number of categories increases, higher rates are allowed and the posterior distribution of site-specific rates has a longer right tail.

BYPASSR is not restricted to the protein coding sequences and I can analyze sep-

site	noncoding sites only	P($r > 1$)	all sites	P($r > 1$)
14	6.089 ±3.1043	0.996	5.520 ±2.6477	0.954
27	4.568 ±2.5352	0.982	4.065 ±2.2431	0.833
193	4.512 ±2.4373	0.981	3.755 ±2.0989	0.800
256	5.238 ±2.9875	0.986	4.842 ±2.6684	0.888
393	5.072 ±2.7804	0.991	4.281 ±2.3529	0.847
408	4.832 ±2.611	0.983	3.891 ±2.2801	0.807

Table 3.5: The sites in the noncoding region with the highest rates obtained by analyzing the noncoding region alone and together with the coding sites. Site numbering corresponds to the alignment with gaps.

arately the noncoding region of 586 sites (537 sites remained after eliminating 49 sites with gaps). For this data set, the estimated tree length is 0.251 ± 0.02 and $\alpha = 0.51 \pm 0.1$. The mean posterior rates of the noncoding sites in the two analyses are comparable, but, when all the sites are considered, the rates at the noncoding sites are scaled to the other site rates. Approximately a fifth of the sites (82 out of 537) have the lowest posterior mean rate of 0.503. These sites correspond to the most conserved sites in the sequences. As table 3.5 shows, the same sites have been identified to have the highest rates. The probability P of having the left tail of the posterior distribution below rate value 1 was different because the values of the rates are smaller in the data set containing all the sites. The fact that I obtained the same sites it is not surprising. The result is in agreement with the simulation results. I demonstrated that the increased number of sites do not affect in the dramatic way the inference of site-specific substitution rates.

As concerns the coding region, I found several sites with high rates and satisfying my criteria for being candidate sites for positive selection. The following candidate sites are identified with high probability P($r > 1$) > 0.95: site 647 (codon 216) with mean posterior rates at codon positions $\overline{r_2} = 5.435$, $\overline{r_1} = 3.056$, $\overline{r_3} = 0.503$, 4292 (1431) $\overline{r_2} = 6.859$, $\overline{r_1} = 0.503$, $\overline{r_3} = 0.503$ and 6962 (2321) $\overline{r_2} = 6.850$, $\overline{r_1} = 1.837$, $\overline{r_3} = 0.503$. None of these codons have been identified by the other parametric methods that analyzed this data set (the coding region of viral DNA only) using dN/dS ratio as the criterion to search for positive selection [29] [39].

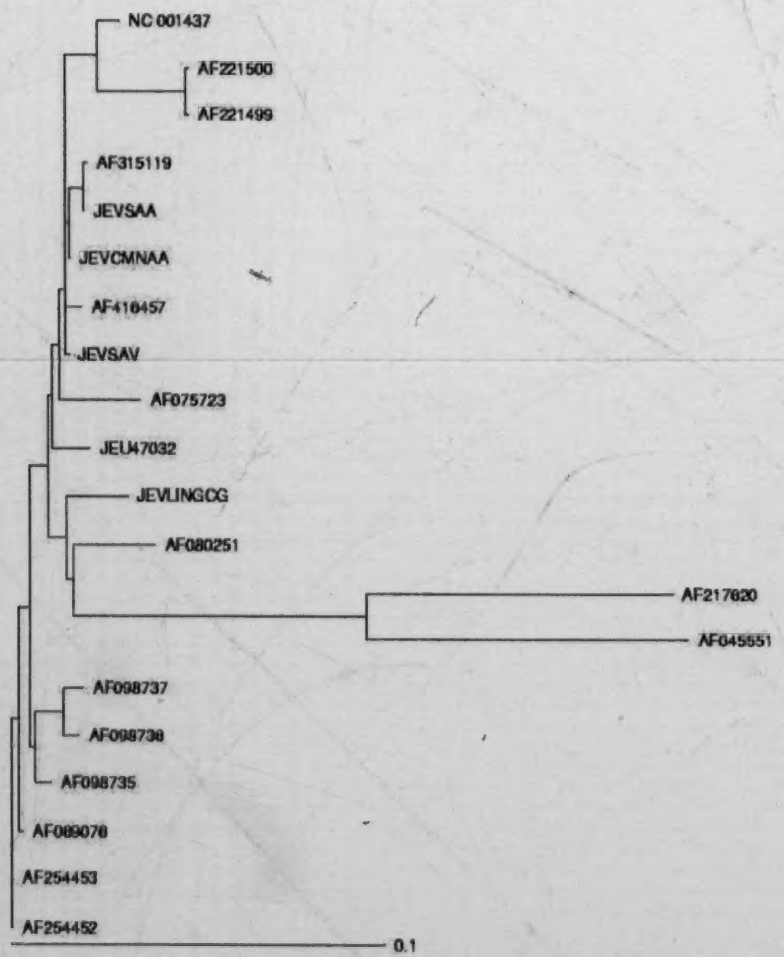


Figure 3.6: Phylogenetic tree of 20 isolates of Japanese Encephalitis complete genome.

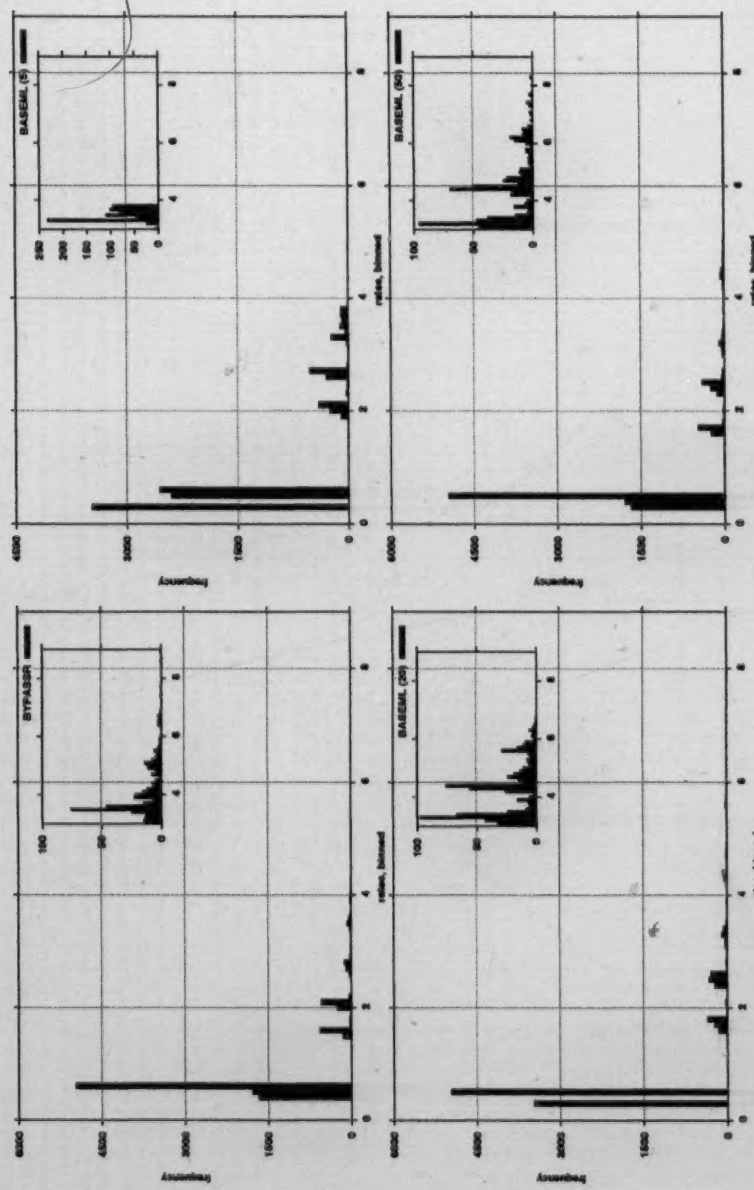


Figure 3.7: Distribution of substitution rates estimates in BYPASSR and BASEML with 5, 20 and 50 categories for Japanese encephalitis virus genome.

3.4 Lyssaviruses Rabies virus Genotype I glycoprotein with 3-prime noncoding region

Lyssaviruses are the causative agents of rabies in mammals. One major cause of virus pathogenity is attributed to the viral surface protein, glycoprotein, which attaches to cell receptors and reacts with host neutralizing antibodies [118] [119]. Glycoprotein comprises a signal peptide (SP), an endodomain, an ectodomain and a transmembrane region (TM). The ectodomain is the largest of them and the most important for viral invasion and escaping host defense mechanisms. The antigenic sites responsible for receptor recognition and membrane fusion, and the sites determining the virulence are all located in the ectodomain. The ectodomain is also the most conserved region of glycoprotein. Because of its important role, positive selection was suspected to act on the entire gene, but this was contradicted by the neutrality tests performed by Badrane and Tordo (2001) [1]. However, they further analyzed the dataset of 55 isolates of rabies glycoprotein in carnivora and chiroptera samples. The pairwise comparison of codon differences along the glycoprotein codons [45] identified regions in SP, TM and Endodomain with dN greater then dS rate, suggesting that positive selection may be present. Wong and Nielsen (2004) [39] also analyzed a part of this data set containing 35 isolates of glycoprotein genotype I together with the downstream noncoding region of 497 nucleotides. They used the approach that combines the analysis of coding DNA with PAML [94] and the implementation of the modified nucleotide substitution model for the noncoding region. Positive selection was not found in the coding or noncoding DNA region.

In spite of solid knowledge about the functions of glycoptotein, the origin and potential functions of the noncoding region located at the 3-prime end of glycoprotein are largely unknown. Considered a pseudogene [120] or transcription regulation factor [121], this noncoding DNA segment has hypervariable sites and a well conserved motif, "GAAAAAAC". Another peculiarity is the maintenance of constant size across all lyssaviruses [1].

Using only the coding sequences of 35 isolates (29 carnivora and 6 chiroptera) of 1572 bp, I constructed the phylogenetic tree with maximum likelihood method available in BASEML (PAML) [94], choosing the GTR model of nucleotide substitution and the gamma distribution of rates with 8 categories. The BYPASSR program was used to analyze the lyssavirus sequences, coding sequences with the downstream noncoding region of 497 bp. The alignment, the same as in the published article of Wong and Nielsen (2004) [39], was kindly provided by Wendy Wong.

Using BYPASSR, I jointly inferred the posterior distribution of site-specific substitution rates, branch lengths, parameters of the GTR model, λ, the parameter of the prior on branch lengths, and α, the parameter of the prior on site-specific rates. I ran the MCMC for 2×10^6 iterations, discarding the first 1×10^6 iterations as burn-in. Inferences were based on 3 independent chains for each run. The effect of using a discrete gamma approximation [50] on the accuracy of maximum likelihood estimates of site-specific

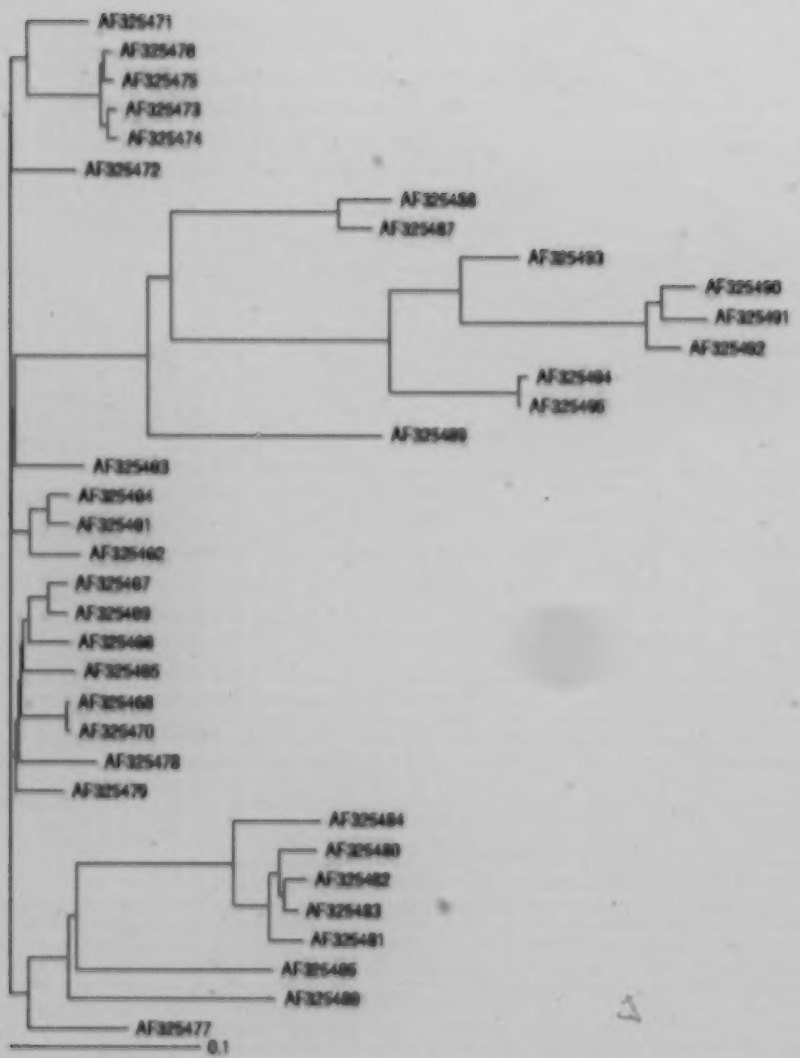


Figure 3.8: Maximum likelihood tree for 35 isolates of lyssavirus.

Param.	BYPASSR (run I)	BYPASSR (run II)	BYPASSR (run III)	BASEML (5 cat.)	BASEML (20 cat.)	BASEML (50 cat.)
α	0.568 ±0.03	0.567 ± 0.03	0.565 ± 0.03	0.536	0.564	0.556
a	1.420 ±0.08	1.425 ± 0.08	1.421 ± 0.08	1.397	1.436	1.431
b	0.089 ±0.01	0.089 ± 0.01	0.089 ± 0.01	0.090	0.088	0.088
c	0.147 ±0.01	0.148 ± 0.01	0.148 ± 0.01	0.147	0.146	0.146
d	0.193 ±0.01	0.193 ± 0.01	0.193 ± 0.01	0.191	0.193	0.192
e	0.028 ±0.01	0.027 ± 0.08	0.027 ± 0.08	0.026	0.026	0.026
TL	2.145 ±0.07	2.130 ± 0.07	2.137 ± 0.07	1.990	2.102	2.138

Table 3.6: Estimates of shape parameter α of gamma prior on site-specific rates, and 5 relative rates from GTR model, a, b, c, d, e, obtained from the mean of the posterior distribution of three independent runs (each with 3 independent chains) of the BYPASSR program (run I, II and III) as well as empirical Bayes estimates from BASEML using a discrete approximation to the gamma distribution with either 5, 20 or 50 rate categories for lyssavirus sequences.

rates, obtained using BASEML was examined by varying the number of rate categories. For purposes of comparison, the same parameters (apart from λ which is not defined for the likelihood method) were estimated by maximum likelihood using BASEML with a GTR model and a discrete gamma approximation with either 5, 20 or 50 rate categories. The estimates from BYPASSR were highly consistent between runs (as judged from a scatterplot of posterior means) and the estimates of $\theta = \{a, b, c, d, e\}$, α and w were also very similar between BYPASSR and BASEML. Table 3.6 presents the estimates of θ and α, obtained from the mean of the marginal posterior densities from 3 BYPASSR runs (each using 3 chains for inferences) and the estimates obtained from BASEML using either 5, 20 or 50 rate categories.

Figure 3.9 shows a scatter plot of the branch length and site-specific rate estimates from BYPASSR (using the mean of the posterior distribution) versus estimates from BASEML with either 5 (panels A, B), 20 (panels C,D), or 50 (panels E,F) rate categories. There is very close agreement between branch length estimates from the two programs even if only 5 rate categories are used (panel A). This agrees with earlier findings [8] that accurate phylogenetic inference can be carried out using a discrete gamma approximation with relatively few rate categories. The three panels on the right site of Figure 3.9 show a scatter plot of site-specific rates (in units of expected numbers of substitutions) inferred using BYPASSR versus BASEML with either 5, 20 or 50 rate categories. With 5 rate categories (panel B), there is a close agreement for rates less than 1 but BASEML appears to overestimate rates at sites with intermediate rates (between 1 and 3) and underestimate rates at sites with high rates (greater than 3). The rate estimates agree more closely with BYPASSR as the number of rate categories increases toward 50, however, even with 50 rate categories (panel D) very high site-specific rates are still systematically underestimated by BASEML.

The mean of the posterior distributions of site-specific substitution rates for all gly-

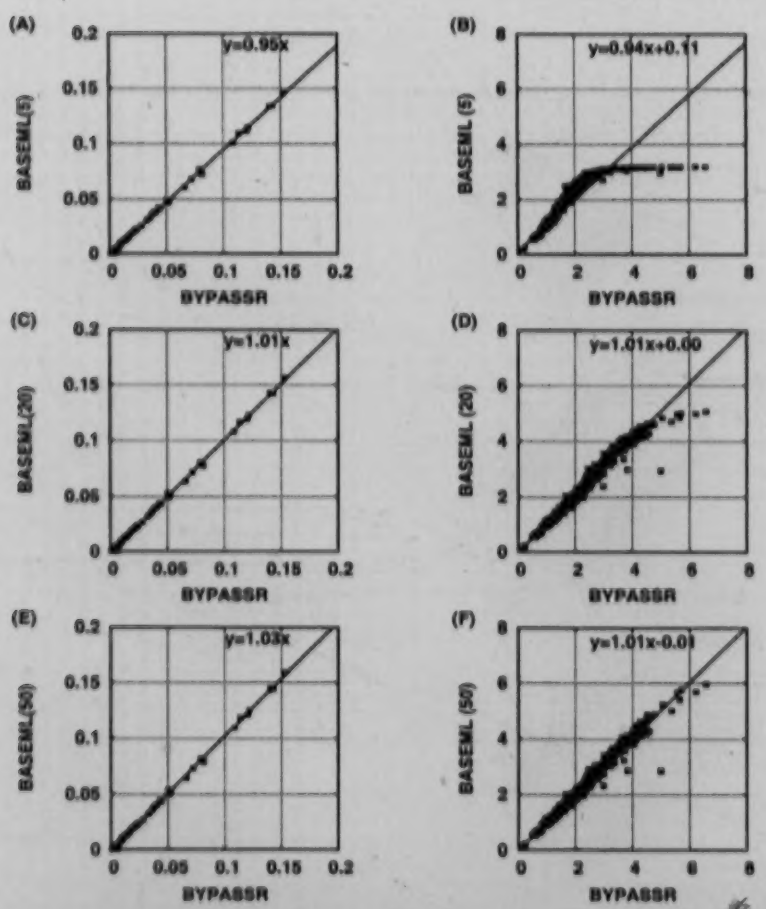


Figure 3.9: Comparison of estimated branch lengths and site-specific rates obtained using BYPASSR and BASEML programs. Left Panels A, C, E plot mean branch lengths from the posterior distribution generated by BYPASSR (horizontal axis) against estimates of branch lengths generated using BASEML with 5, 20 and 50 rate categories (vertical axis). Right Panels B, D and F plot the mean site-specific rates from the posterior distribution generated by BYPASSR (horizontal axis) against estimates of site-specific rates generated using BASEML (vertical axis) with either 5, 20 or 50 rate categories, respectively.

	nucl	$\overline{r_1} < \overline{r_2} > \overline{r_3}$	mode r_2	$P(r_2 > 1) > 0.8$
1	635	2.103	1.6	0.917
nucl	$\overline{r_1} > \overline{r_3}$	mode r_1	$P(r_1 > 1) > 0.8$	
1	361	2.47	1.4	0.9
2	523	2.67	2.4	0.967
3	604	2.02	1	0.822
4	700	2.36	1.6	0.934
5	736	2.28	1.4	0.912
6	883	1.88	1.6	0.834
7	907	3.22	2.2	0.994
8	1252	2.77	2.6	0.963
9	1336	2.89	2.2	0.985

Table 3.7: The ectodomain candidate sites to evolve under positive selection.

coprotein coding and 3-prime noncoding region are shown Figure 3.10. Ectodomain shows the trend that is typical for coding regions with third codon positions having the highest substitution rates and second codon positions the lowest, with rates at first codon positions intermediate between these two extremes. Most of the ectodomain sites, including the 16 cysteine sites (17, 43, 54, 80, 113, 178, 188, 208, 226, 242, 247, 271, 302, 363, 370, 479) have substitution rates at the first and second codon positions less than 0.25. This is the expected pattern for negative selection acting at the level of the amino acid sequence. My results confirm the previous observation [1] of a high degree of conservation of the ectodomain sites. However, I identified 10 nucleotide sites that satisfy my positive selection criteria. The probability of $r > 1$ for these sites is greater than 0.8 (Table 3.7).

The SP and TM domains (Fig 3.11 panel A) also have high rate variation among sites with a few sites possibly undergoing positive selection. Endodomain (Fig 3.11 panel B) is atypical with rates at second codon positions exceeding those of first codon positions for a large proportion of sites (Fig 3.11 panel B). This may indicate positive selection operating in this domain. None of these potentially positively selected sites were identified with the maximum likelihood method as implemented in CODEML (PAML) [94].

Fig 3.11 panel C shows the posterior means of substitution rates at glycoprotein 3-prime flanking noncoding region. High substitution rates ($\overline{r} > 3$) are observed at 33 sites out of 497 and the majority of rates exceed the overall mean of rates 1. A few blocks of sites have very low substitution rates revealing the conserved areas across the sampled sequences. The conserved motif "GAAAAAAAC" located at sites 2036-2045 is among the sites with the lowest substitution rates in the whole dataset, therefore good candidate sites for evidence of purifying selection.

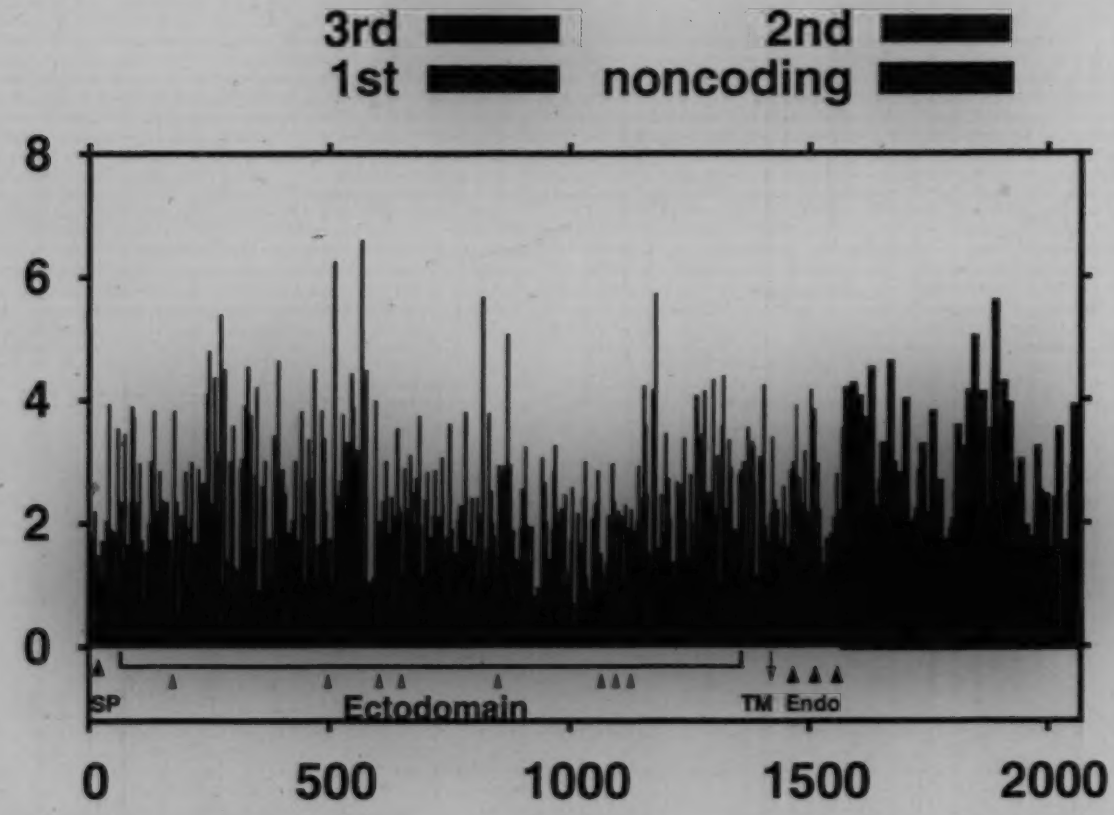


Figure 3.10: Mean posterior rate at codon positions of glycoprotein gene and at the sites located in the 3-prime noncoding region. The approximate location along the gene of the signal peptide (SP), endodomain, ectodomain, transmembrane domain (TM), antigenic sites and Western Blot positive epitope (open arrow heads) is shown as in [1].

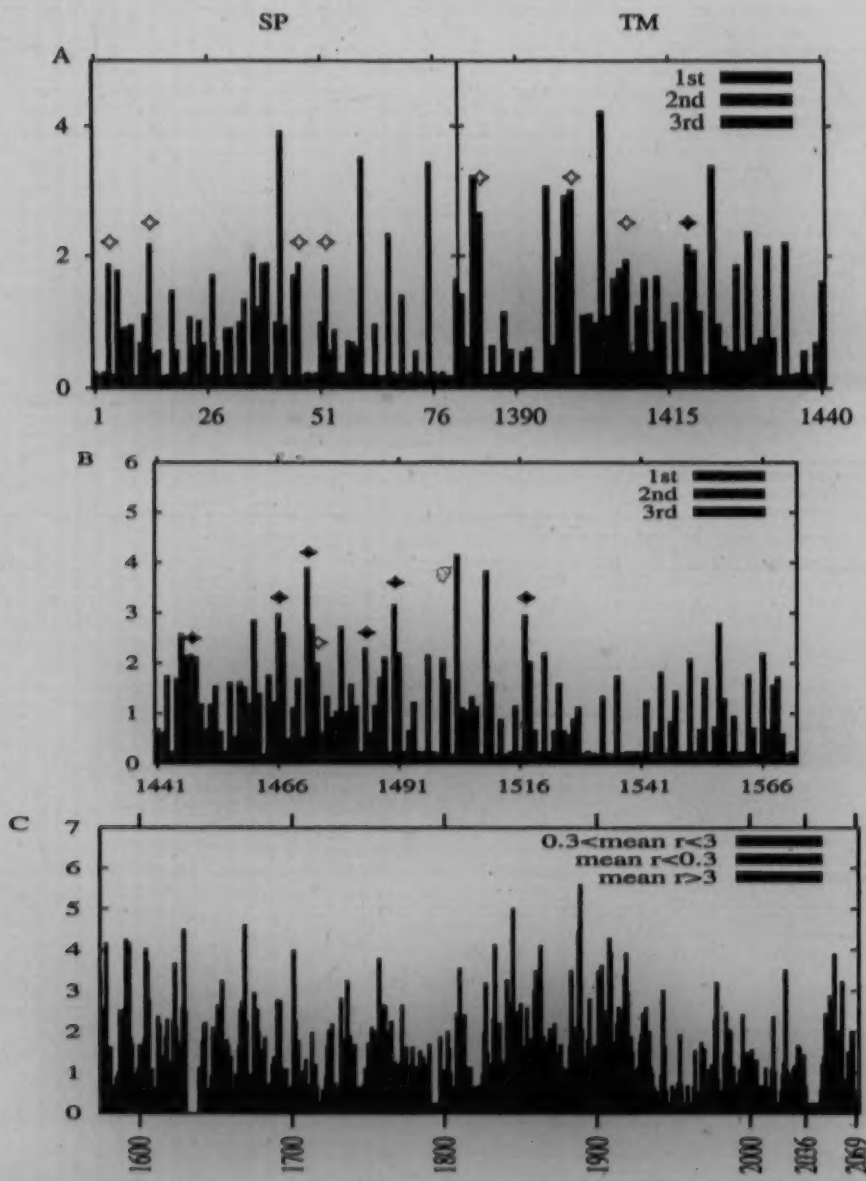


Figure 3.11: Mean posterior distributions at first, second and third codon positions of signal peptide and transmembrane domains (A) and endodomain (B). The sites located at the second position or first codon position that are good candidates to evolve under positive selection are marked with ✦ and ✧, respectively. Panel C shows the mean posterior rates for sites in the 3-prime noncoding region.

3.5 Mammalian cytochrome b

Cytochrome *b* (*cytb*) is an extensively studied gene because of its important metabolic functions in mitochondria. The protein is approximately 400 amino acids divided into three major functional domains: an intermembrane domain composed of 7 loops (ab, bc, cd, de, ef, fg and gh), transmembrane domains formed by 8 α helices (A to H), a matrix domain composed of the N and C termini and tree loops entering the mitochondrial matrix. Comparative studies of sequence variation have found that the transmembrane domain evolves the fastest among the domains, followed by the matrix domain, with relatively few conserved residues [122]. The slowest evolving domain was found to be the intermembrane domain [122]. Because of its sequence conservation among closely related species, cytochrome *b* gene gives reliable alignments and is usually used for establishing relationships among eukaryotic organisms [123] [124] [125]. For example, a set of 28 sequences of primate species was searched for positively selected sites and an average ω across sites of 0.04 was found, suggesting a high degree of gene conservation [126]. However, the study of quantitative biochemical properties of *cytb* in a limited number of cetacean and artiodactyl species identified potential positively selected sites in great number [127]. The transmembrane domains host the majority of the 90 candidate codons found by McClellan *et al.* (2005) [127]. Two loops of the matrix domains, three intermembrane domains and the N-terminus matrix domain contain at least 3 residues possibly affected by positive selection.

A total number of 688 sequences is available on the cytochrome *b* server [124]. The sequences of length varying between 1005 and 1140 nucleotides were aligned with the default options in *clustalw* software [128] that also generated the neighbor joining (pairwise distance method) tree of length 40.132. The first and the last 44 codons are eliminated because of multiple gaps across sequences. After elimination of the sites with gaps, only 694 sites remained for inference. The chain ran for an 91×10^6 iterations followed by other 20×10^6 from which 2000 samples for statistical analysis are collected. The total running time was approximately 6 weeks. The burn-in had to be restarted multiple times with the initial values saved at the end of the previous runs. Samples of the α, λ, tree length and log likelihood of the tree were recorded. Fig.3.12 shows how they changed along the runs. I also calculated the correlation matrix to verify if the site rates are somehow correlated. Among the total of 694 sites used in the analysis 120 sites, all with rates greater than 1, had the correlation coefficient between 0.40 and 0.43. Considering that every 10,000th sample was chosen to do the statistics, the higher than expected correlation coefficient (i.e. no correlation in a model built under site independency assumption) is unlikely to be caused by too frequent sampling. The correlation matrix for the first 100 sites and for the last 100 sites are shown in Fig. 3.13. For this data set posterior means of the parameters of interest are: $\alpha = 0.29 \pm 0.02$, tree length = 62.61 ± 2.1, $\lambda = 22.04 \pm 0.95$, and the relative rates of the GTR $a = 2.27, b = 0.23, c = 0.07, d = 0.27, e = 0.06$. As the simulation results have shown, increasing the number of taxa determined the posterior distribution of rates to be sharper and with a clearly defined mode. My expectations are met and the rates are inferred with great

confidence.

Using my criteria to select individual nucleotide sites that might evolve under positive selection in protein coding sequences, I found many sites located at the second or first codon position with the substitution rate greater than the substitution rate at the site occupying the third position in a codon. The sites with the highest rates in the dataset, satisfying my criteria also, with the probability of having more than 95% of the posterior distribution above $r = 3$ are plotted in Fig. 3.15. The sites that have $\overline{r_1}$ or $\overline{r_2}$ between 2 and 3 and still are greater than the rate at the third codon position are shown in Fig 3.16. Many of the sites found by us to be good candidates for positive selection to act upon are not identified in previous analyzes of cytochrome *b* in much smaller datasets. The lack of methods to analyze such a large number of sequences missed many sites that seem to be important for evolution of this gene. My results strongly suggest that positive selection might be more prevalent in cytochrome *b* that previously thought.

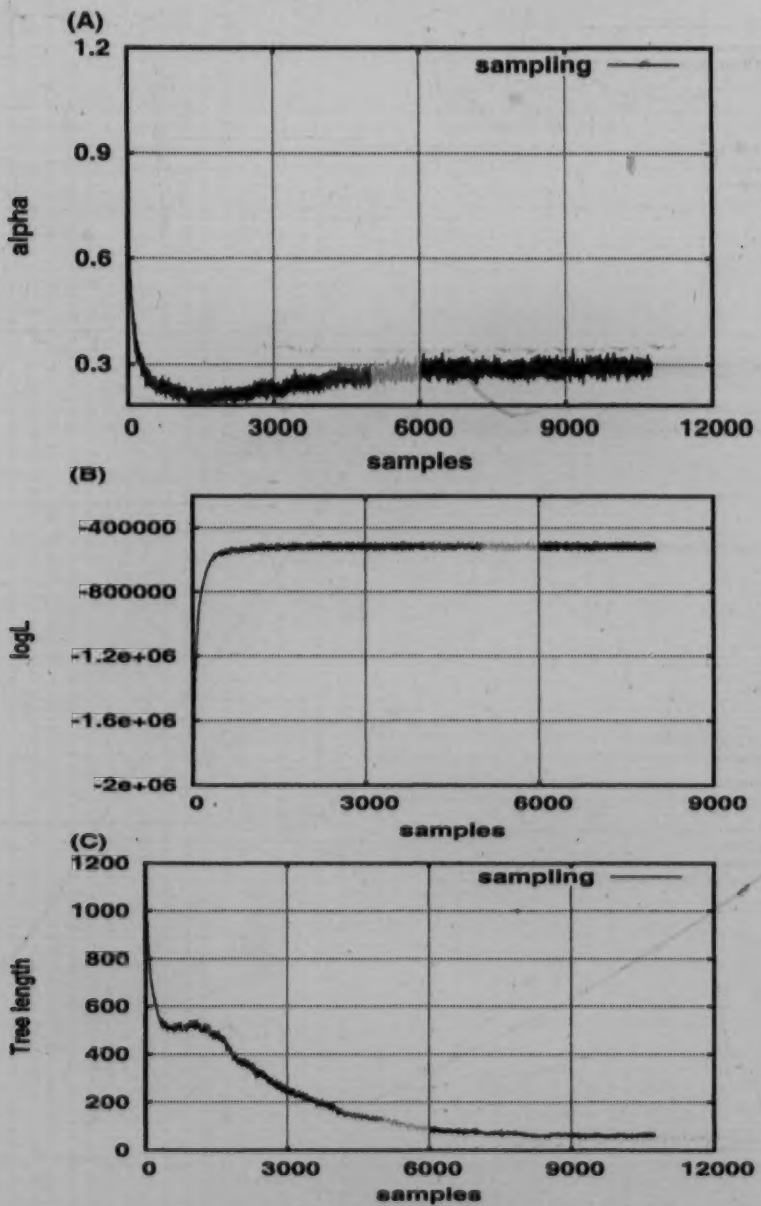


Figure 3.12: Chain convergence during the 9 burn-in runs and sampling for α Panel (A), log L tree (B) and tree length (C) for cytochrome *b* dataset .

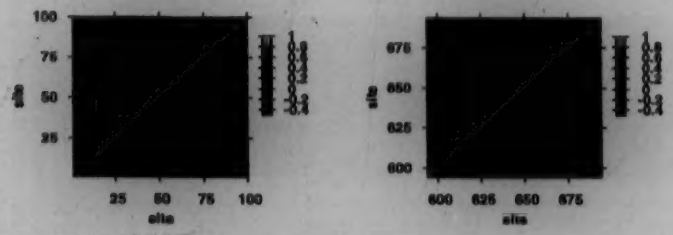


Figure 3.13: Correlation matrix for the first 100 sites (left panel) and for the last 100 sites (right panel).

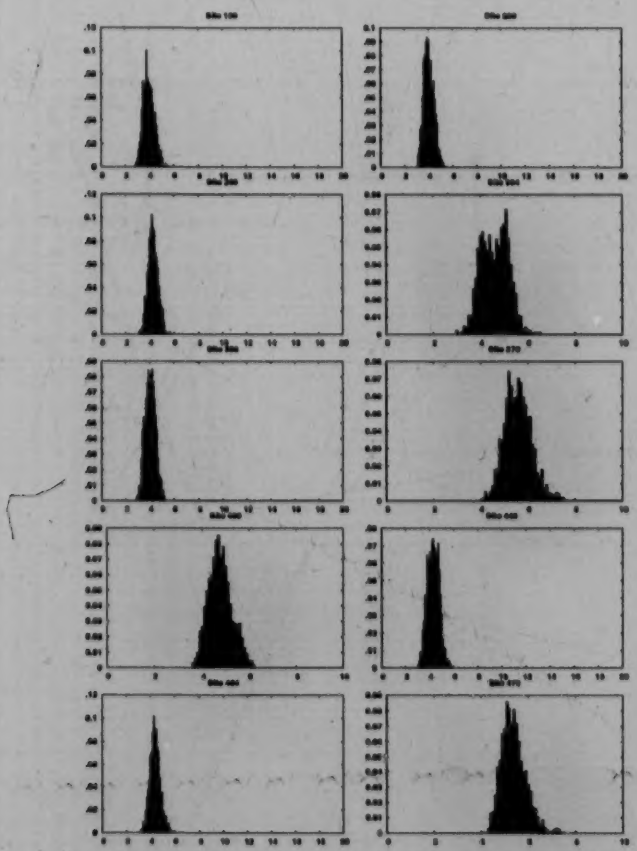


Figure 3.14: Posterior distribution at some sites with high rates.

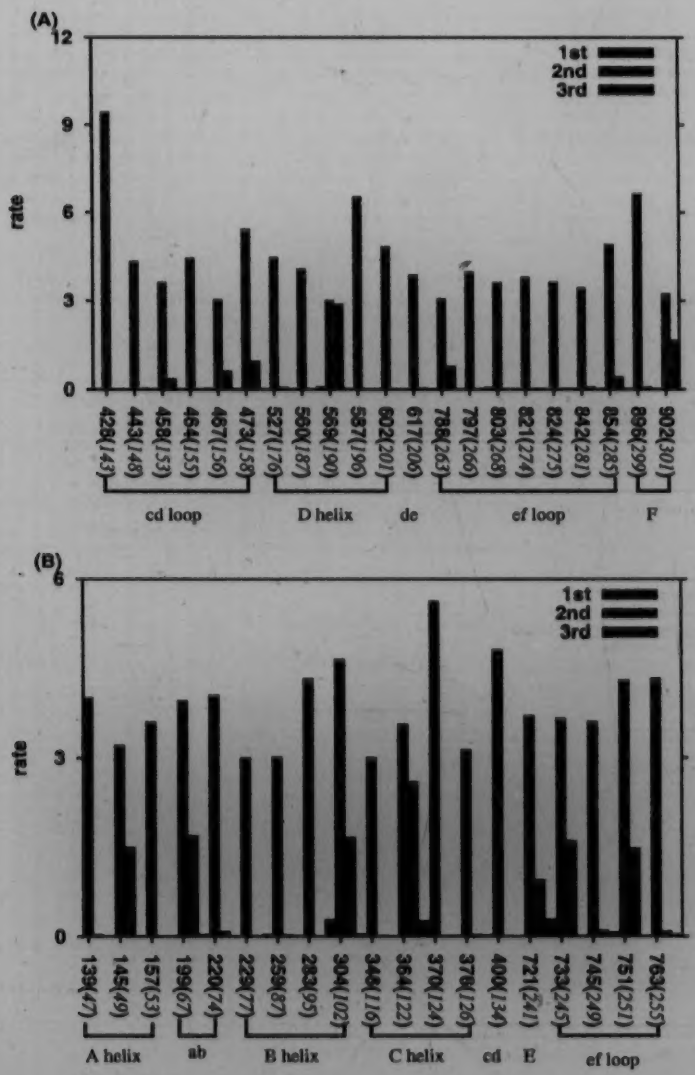


Figure 3.15: Mean posterior substitution rates at candidate sites for positive selection in cytochrome *b* gene. Panel A shows the codons that have $\overline{r_3} < \overline{r_2} > \overline{r_1}$ and $\overline{r_2} > 3$ with probability >0.95 in italics. The codons with $\overline{r_1} > \overline{r_3}$ and $r_1 > 3$ with probability >0.95 (italics) are in the bottom panel.

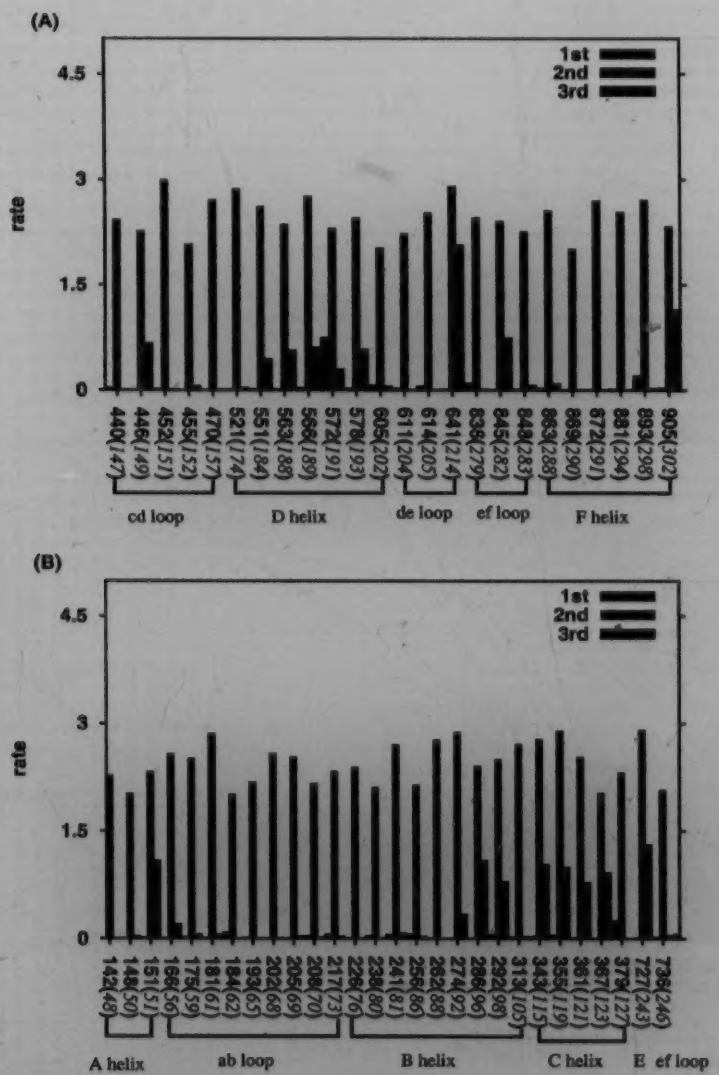


Figure 3.16: Mean posterior substitution rates at candidate sites for positive selection in cytochrome *b* gene. Top panel show the codons that have $\overline{r_3} < \overline{r_2} > \overline{r_1}$ and $2 < \overline{r_2} < 3$ with probability >0.95 in italics. The codons with $\overline{r_1} > \overline{r_3}$ and $2 < \overline{r_1} < 3$ with probability >0.95 (italics) are in the bottom panel.

3.6 Reconstructing the ancestral sequence of EDN and ECP genes

A gene duplication event is an important cause of evolution because, after duplication, the two genes may follow independent evolutionary paths giving chance to novel functions to occur. This was proposed to be the history of the eosinophil-derived neurotoxins, EDN, or RNase 2 and ECP, or RNase 3, members of the ribonuclease superfamily. While only a single copy of the gene EDN/ECP is found in New World monkey and prosimians (EDN), the Old World monkeys and hominids have both. The antiviral ability of EDN gene, digesting the RNA of retroviruses, is substantially enhanced in Old World monkeys and hominids after the duplication event. The same lineages benefit from the antibacterial and antiparasitic activity of the ECN gene, through a mechanism that is RNase independent. The functions of the ancestral protein predating the duplication event and the amino acids substitutions causing the improvement of the antiviral activity of EDN in one descendant lineage are intriguing evolutionary questions. A "paleomolecular biochemical" approach was taken to reconstruct the ancestral protein of EDN and ECP gene [2]. First, the sequences at the internal nodes are statistically inferred, followed by experimental mutations to potentially relevant sites in EDN gene. The study is based on 13 EDN sequences and 5 ECP sequences corresponding to primates. The phylogenetic tree was obtained from the amino acid alignment and the inference of the ancestral nucleotide states was done with distance-based Bayesian method (branch lengths estimated with least square method and the ancestral amino acids inferred with empirical Bayes method) [71] and parsimony (branch lengths have no relevance and the number of changes between two states is minimum) [53] methods. The authors are focused on reconstructing the ancestral sequences at the node A and B (Fig 3.17) for which they obtained an overall posterior probability greater than 0.93 for both nodes.

The important sites for protein function are the three major catalytic residues (15H, 38H and 125H) and the structural cysteine residues located at 23, 37, 55, 62, 71, 83, 93, 108 (site numbers corresponds to alignment after eliminating sites 87, 88, 90 and 117 because of missing data). Statistical reconstruction of the proteins at node A and B found 9 amino acids substitutions between the two. Conclusions of earlier studies and mapping these amino acid sites on the crystal structure of human EDN, reduced the number of important sites to two. In 3D view, Ser-64 and Arg-132 are in close vicinity and close to His-125H, a known catalytic site. After synthesis of the proteins with and without substitutions at these sites, the RNase activity was measured through biochemical experiments and a 13-fold enhancement was found when both substitutions were present. The authors further conclude that both substitutions at amino acid site 64 (Arg$-$ $>$ Ser) and 128 (Thr $-$ $>$ Arg) at node B are found to be necessary to explain the potent activity of EDN in hominids lineage.

The uniformization technique and data augmentation approach allow the inference of nucleotides at the internal nodes without any additional modification of the code. This feature of BYPASSR is used to reconstruct the DNA sequences at all internal

nodes. The nucleotides having the highest posterior probabilities at the two nodes of interest obtained with BYPASSR are also compared with the marginal reconstruction of ancestral states at node A and B in BASEML (PAML) [94]. I used the published tree topology and the reverse translated sequences of 130 amino acids (*backtranseq* option in *EMBOSS* [129]). Using the GTR model of DNA substitution and the molecular clock assumption (constant rate of evolution among primates), a BYPASSR run achieved convergence in less than an hour (1 million iterations) and 5000 samples for each of the parameters of interest (branch lengths, substitution rates, parameters of the GTR substitution model and nucleotides at the internal nodes) were collected. The posterior probability of a nucleotide is given by its occurrence frequency in the samples collected at that node and that site. For example, if a G nucleotide at a site and a node is present in 1500 out of 2000 samples that gives the posterior probability of 0.75. The alignment in Fig 3.18 shows the comparison between the amino acids at the nodes of interest obtained previously and the sequences inferred with BYPASSR and BASEML (with different numbers of categories for gamma distribution). The overall posterior probabilities in BYPASSR at node A and B are 0.95 and 0.97, respectively. As expected, the accuracy of reconstruction is related to the distance of the inferred sequence from the extant sequences. Being located closer to the root, the hypothetical protein at node A has a smaller overall posterior probability. The 3 amino acids with posterior probability less than 0.5 in the published study, have low posterior probability in BYPASSR also: 21Q [C(0.986) C(0.348) G(0.496)], 66K [A(0.962) C(0.68) G(0.594)] and 97Q [G(0.438) A(0.959) C(0.884)]. The tables 3.9 and 3.8 show the nucleotides with their posterior probabilities for the sites at nodes A and B that resulted in different amino acids than previously obtained. The posterior probabilities of BASEML marginal reconstruction using different number of categories for gamma distribution approximation are also tabulated. Inference of ancestral character states does not appear to be influenced by the number of categories in gamma. In general, BYPASSR obtained slightly lower posterior probabilities than BASEML. The explanation resides in the manner by which Bayesian and maximum likelihood methods deal with uncertainty. Maximum likelihood fixes the parameters to the values that maximizes the tree likelihood without considering any uncertainty in the estimates, while the Bayesian approach uses the whole distribution of the parameters not only their "best" values. However, the only differences between the two model-based methods are explained by very low posterior probabilities at those nucleotide sites. Among the 12 amino acids found different by BYPASSR or BASEML at node A, sites 20P and 64S have high posterior probabilities in both. Node B reconstruction differs at 6 sites, with high confidence for 45L. As the study focuses on site 64 and 128 being essential in explaining the functions of EDN and ECP gene, some of their interpretations might be misleading if the results obtained with BYPASSR and BASEML are correct. The disagreement between the results may be related to the known problems of parsimony or distance methods in the reconstruction of ancestral states, usually explained by the simplicity of assumptions about evolutionary processes used in these methods.

amino acid site	bypassr	baseml 50 cat.	baseml 20 cat.	baseml 5 cat.
7	C(0.827)	C(0.826)	C(0.829)	C(0.852)
8	C(0.403)	G(0.386)	G(0.385)	G(0.364)
	G(0.385)			
	A(0.211)			
9	C(0.947)	C(0.929)	C(0.930)	C(0.935)
133	C(0.956)	C(0.955)	C(0.955)	C(0.958)
134	T(0.936)	T(0.941)	T(0.941)	T(0.942)
135	G(0.998)	G(0.997)	G(0.997)	G(0.997)
148	A(0.527)	A(0.539)	A(0.539)	A(0.541)
	G(0.466)			
149	A(1.000)	A(1.000)	A(1.000)	A(1.000)
150	C(1.000)	C(1.000)	C(1.000)	C(1.000)
196	A(0.996)	A(0.998)	A(0.998)	A(0.998)
197	C(0.654)	C(0.711)	C(0.710)	C(0.702)
	A(0.221)			
198	G(0.592)	G(0.546)	G(0.545)	G(0.541)
	C(0.402)			
199	A(1.000)	A(1.000)	A(1.000)	A(1.000)
200	C(0.521)	G(0.523)	G(0.523)	G(0.519)
	G(0.474)			
201	C(1.000)	C(1.000)	C(1.000)	C(1.000)
286	G(0.551)	A(0.510)	A(0.510)	A(0.512)
	A(0.443)			
287	C(0.996)	C(0.997)	C(0.996)	C(0.996)
288	C(1.000)	C(1.000)	C(1.000)	C(1.000)

Table 3.8: Posterior probabilities obtained with BYPASSR and empirical Bayes estimates of BASEML (with 5, 20 and 50 categories) of the nucleotides at ancestral node B that resulted in different amino acids that indicated in Zhang and Rosenberg study.

amino acid site	bypassr	baseml 50 cat.	baseml 20 cat.	baseml 5 cat.
7	C(0.931)	C(0.937)	C(0.941)	C(0.959)
8	C(0.401)	C(0.410)	C(0.410)	C(0.413)
	A(0.340)			
	G(0.254)			
9	C(0.755)	C(0.819)	C(0.819)	C(0.829)
34	G(0.840)	G(0.943)	G(0.943)	G(0.939)
35	G(0.430)	G(0.420)	G(0.419)	G(0.416)
	A(0.316)			
	C(0.252)			
36	C(0.812)	C(0.831)	C(0.829)	C(0.822)
58	C(0.889)	C(0.902)	C(0.902)	C(0.894)
59	C(0.884)	C(0.886)	C(0.886)	C(0.879)
60	C(0.999)	C(1.000)	C(1.000)	C(1.000)
64	A(0.605)	A(0.550)	A(0.549)	A(0.549)
	C(0.348)			
65	A(0.600)	A(0.515)	A(0.515)	A(0.521)
	G(0.388)			
66	G(0.984)	G(0.993)	G(0.993)	G(0.992)
133	C(0.874)	C(0.871)	C(0.870)	C(0.872)
134	A(0.554)	A(0.510)	A(0.510)	A(0.516)
	G(0.222)			
	T(0.178)			
135	G(0.841)	G(0.944)	G(0.944)	G(0.940)
148	G(0.531)	A(0.533)	A(0.534)	A(0.536)
	A(0.449)			
149	A(0.996)	A(0.998)	A(0.998)	A(0.998)
150	C(0.999)	C(1.000)	C(1.000)	C(1.000)
190	A(0.944)	A(0.956)	A(0.961)	A(0.978)
191	G(0.962)	G(0.972)	G(0.971)	G(0.971)
192	C(0.733)	C(0.730)	C(0.733)	C(0.744)
	G(0.234)			
196	A(0.841)	A(0.949)	A(0.949)	A(0.946)
197	C(0.669)	C(0.701)	C(0.699)	C(0.690)
198	G(0.508)	G(0.536)	G(0.535)	G(0.531)
199	A(0.999)	A(1.000)	A(1.000)	A(1.000)
200	G(0.531)	G(0.521)	G(0.520)	G(0.517)
	A(0.455)			
201	C(0.999)	C(1.000)	C(1.000)	C(1.000)
223	G(0.516)	A(0.527)	A(0.527)	A(0.524)
	A(0.471)			
224	G(0.999)	G(1.000)	G(1.000)	G(1.000)

225	C(0.811)	C(0.810)	C(0.810)	C(0.811)
286	A(0.512)	A(0.513)	A(0.513)	A(0.515)
	G(0.473)			
287	C(0.828)	C(0.934)	C(0.934)	C(0.930)
288	C(0.999)	C(1.000)	C(1.000)	C(1.000)
289	G(0.363)	G(0.417)	G(0.417)	G(0.413)
	A(0.327)			
	C(0.310)			
290	A(0.837)	A(0.949)	A(0.949)	A(0.946)
291	C(0.88)	C(0.886)	C(0.886)	C(0.879)

Table 3.9: Posterior probabilities obtained with BYPASSR and empirical Bayes estimates of BASEML (with 5, 20 and 50 categories) of the nucleotides at ancestral node A that resulted in different amino acids that are indicated in the Zhang and Rosenberg article.

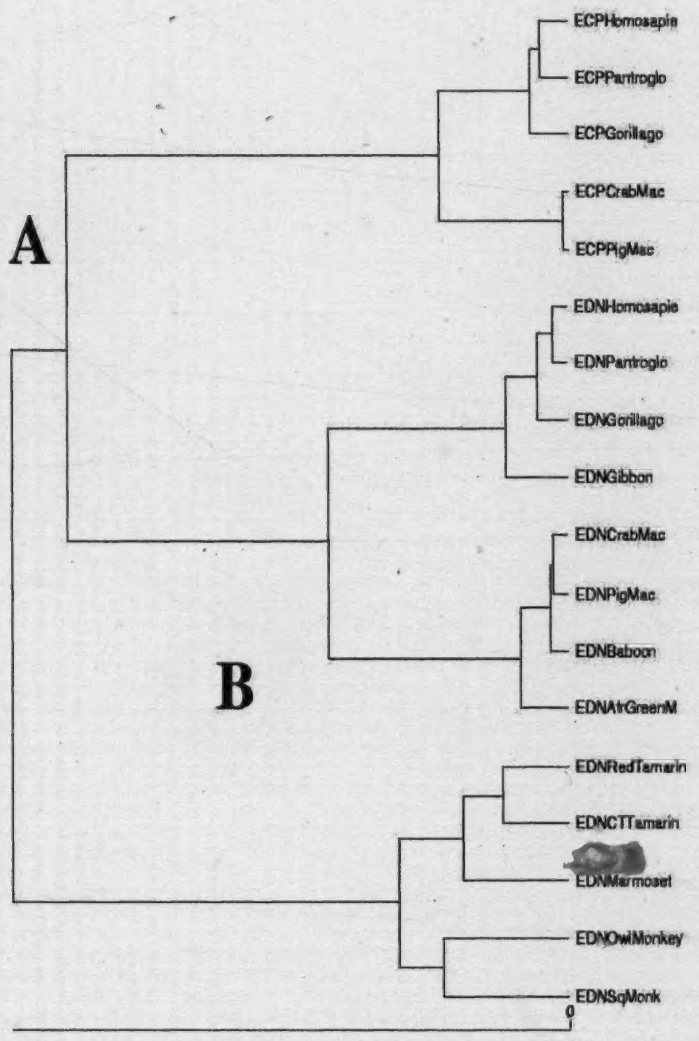


Figure 3.17: Phylogenetic tree of the EDN/ECP gene for 18 primate sequences. [2].

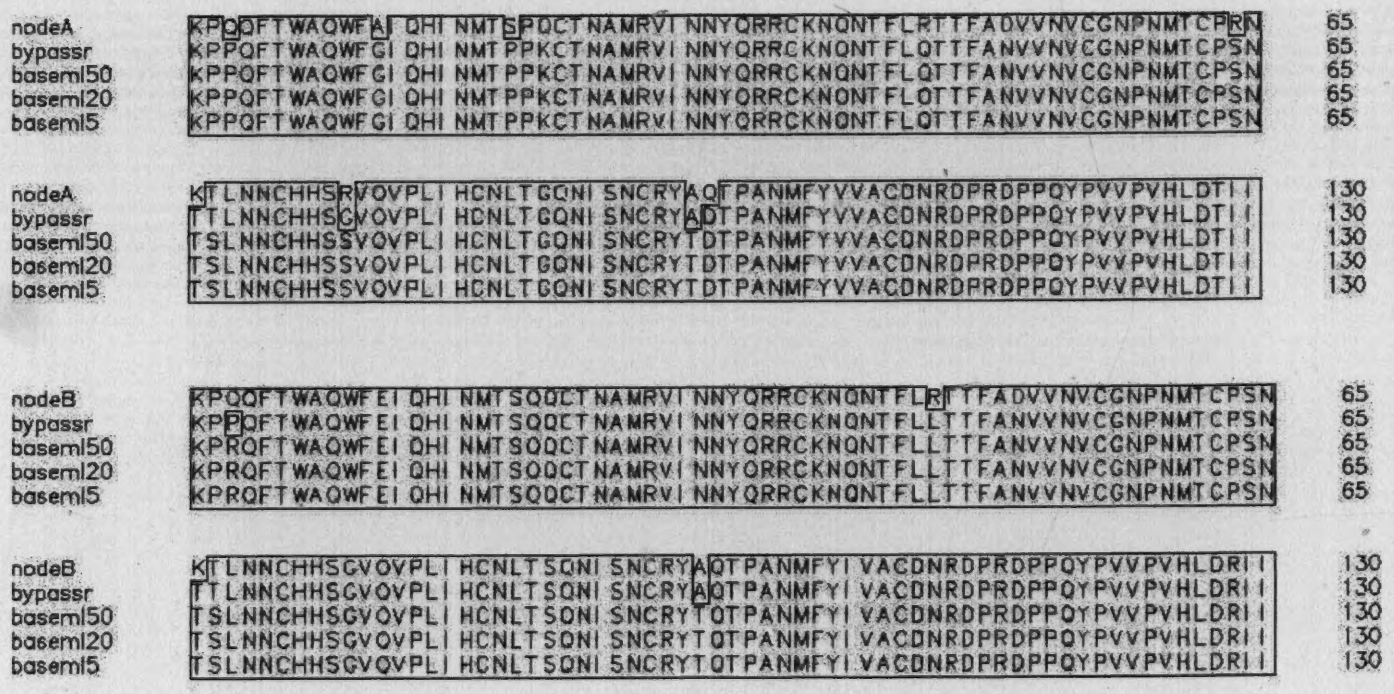


Figure 3.18: Protein reconstruction at internal nodes A and B as inferred with parsimony, Bayesian and maximum likelihood methods

The posterior mean of α for this dataset is 0.679±0.15 indicating heterogeneity of substitution rates among sites. There are 249 out of 390 sites (63%) with substitution rate < 0.6 and only 8 sites with rate greater than 3. Among the sites with lowest rates are the ones corresponding to amino acids important for functional and structural properties of EDN/ECP gene (15H, 38H, 125H, 23C, 37C, 55C, 62C, 71C, 83C, 93C, 108C). BYPASSR identified correctly the sites known to be conserved. It is worth comparing the sites identified by BYPASSR as the candidates for positive selection evolution with the codons found by CODEML to have $\omega > 1$. Mean posterior substitution rates for all the sites of EDN/ECP gene are plotted in Fig 3.19. The sites with the mean posterior rate at the second or first position greater than the other two rates of the codon and probability of $r > 1$ greater than 0.75 are my candidate sites. The majority of them belong to codons identified by CODEML to be under position selection (marked with ❷). BYPASSR found 6 additional sites that might suggest positive selection is operating on them (marked with ❶). Although, there is no direct correspondence between the two methods, I observed a condition when CODEML fails to identify the codons in which one of the sites might undergo selection according to my method. As I presented in the introductory section, codon-based models are based on a greater number of assumptions than DNA substitution models. One such assumption considers that nonsynonymous substitutions are caused mostly by changes at the first nucleotide position in the codon. The reasoning for this assumption is given by the similar chemical properties of the amino acid resulting from such a nucleotide change. CODEML excludes the nucleotide substitution at the second position just because they are typically radical changes. This explains why CODEML did not find the codons with high substitution rates at the second codon position found by BYPASSR in many data sets analyzed by us. On the other hand, in dealing with nucleotide substitution models, BYPASSR is relating the substitution rate at a codon position to the other two rates (i.e. substitution rate from a codon to another is not defined). This explains why CODEML identified 7 additional codons to evolve under positive selection (❸).

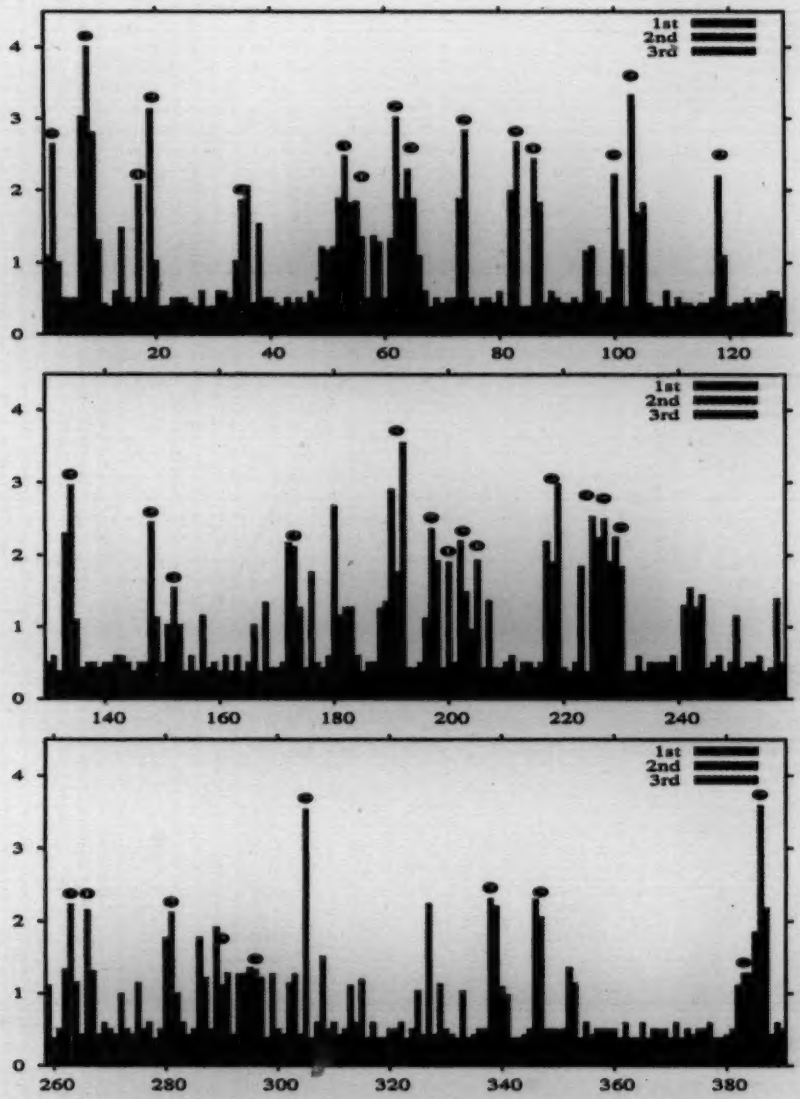


Figure 3.19: Mean posterior substitution rates at first, second, and third codon positions of the EDN/ECP gene. The possible nucleotide sites found by BYPASSR to evolve under positive selection (❶). The codons found by CODEML to be positively selected and also met criteria of containing a positively selected nucleotide site in BYPASSR (❷); The extra codons found by CODEML (❸).

Chapter 4

Modeling uncertainty in the fossil data

The model described in this chapter was presented in June 2006 at *Evolution* with the talk entitled "A Bayesian Method For Modeling Degradation Of Ancient DNA", Stony Brook University, NY.

4.1 Model description

The miscoding lesions generated during amplification procedures of ancient DNA template are characterized by four types of substitutions with two phenotypic outcomes: A $\longrightarrow$ G / T $\longrightarrow$ C and C $\longrightarrow$ T / G $\longrightarrow$ A [81]. Miscoding lesions were detected in samples of 4-year-old dried tissues [79], as well as tissues as old as thousands of years [130]. As the accumulation of substitutions is not a strict function of time, the generation of miscoding lesions cannot be modeled in the same way as the substitution process on the branches of a phylogenetic tree. Instead, a discrete Markov process in which the 4 possible substitutions are allowed with a small rate is a simple and straightforward way to describe the process.

$$\mathbf{D} = \begin{pmatrix} p & q & z & z \\ q & p & z & z \\ z & z & p & q \\ z & z & q & p \end{pmatrix},$$

where each line has to sum to 1 and the rows and columns represent A, T, C and G.

Most of the nucleotides are expected to not be affected by degradation and this is manifest as a value of p close to 1. Hofreiter *et al.* (2001) [82] found that the chemical treatment of the DNA template with uracil N-glycosylase reduces the number of substitutions attributed to the amplification errors to 0.1% of the nucleotides in ancient DNA. However, the true number of these type of substitutions is not known. The amplification errors are described by the parameter q in our model, while the unlikely transversions are represented by the parameter $z \approx 0$. In our approach, the parameters p, q and z are not fixed to particular values. Instead, we consider data informative enough to decide if

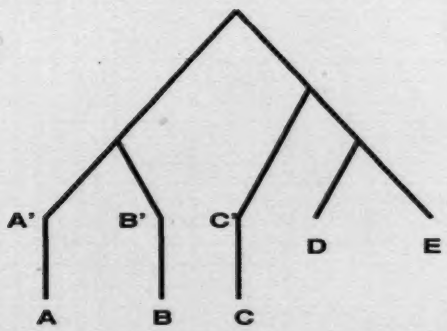


Figure 4.1: Phylogenetic tree under degradation model.

any of the degradation transitions are likely to occur at a few sites of some of the extinct sequences, independent of all the circumstances in which degradation occurred.

If the data does not support degradation, the matrix will take the form of the identity matrix, with $p = 1$ and $q = z = 0$. In other words, we can think of the degradation process as a substitution process happening on an edge that connects the sequence extracted from the ancient DNA with a hypothetical sequence that existed in the past. In Fig. 4.1, this is represented by moving the sequences from the nodes with ancient DNA to the tips along the "degradation" edges, while the nucleotides at nodes A', B', C' are obtained according to their probabilities in the stochastic matrix above. Formally written, for U ancient DNA sequences, each of length n, the augmented likelihood becomes

$$f(\mathbf{M}, \mathbf{x}, \mathbf{x}^{-}, \mathbf{x}^{\dagger} | \mathbf{r}, \tau, \pi, \theta, \mathbf{D}) = \prod_{m=1}^{n} \prod_{l=1}^{2s-3} \prod_{u=1}^{U} \prod_{v=1}^{n} f(\mathbf{x}_m, \mathbf{x}_m^{-} | \theta, M_{lm}, r_m, w_l, \pi, T) \Pr(M_{lm} | r_m, w_l) \Pr(x_{uv}^{\dagger} | D_{uv}), \quad (4.1)$$

where u is the hypothetical sequence and $x_{uv}^{\dagger}$ is a nucleotide at the hypothetical sequence u at site v.

4.2 Implementation

In our model, the degradation process is a time independent process and the age of the ancient DNA is irrelevant. BYPASSR was first modified to distinguish between ancient and current DNA and to allow the addition of the hypothetical nodes by moving the fossils sequences to a level below. The nucleotides at the hypothetical nodes become random variables in the chain together with the degradation parameters p, q and z. The remaining parameters in the chain do not interfere with the degradation process parameters as the nucleotides at the hypothetical nodes are considered fixed.

Modifying degradation parameters

A sampling distribution has to be specified for the degradation parameters. Parameters p, q and z represent probabilities and a prior that constraints the rows to sum to 1 is needed. The Dirichlet distribution is a commonly used conjugate prior. The probability density function of the Dirichlet distribution is the function of a vector of 3 parameters $\mathbf{x} = (x_1 = p, x_2 = q, x_3 = z)$

$$f(x|a_0) = \frac{1}{\mathbf{B}(a_0)} \prod_{i=1}^{3} x_i^{a_i - 1}, \tag{4.2}$$

where $a_0 = (a_1, a_2, a_3)$ is the parameter vector with $a_i \geq 0$ and $\mathbf{B}$ is the normalizing constant

$$\mathbf{B}(\alpha_0) = \frac{\prod_{i=1}^{3} \Gamma(a_i)}{\Gamma(\sum_{i=1}^{3} a_i)}. \tag{4.3}$$

The marginal means and variances of the distribution are a_i/a_0 and $a_i(a_0 - a_i)/a_0^2(a_0 + 1)$, respectively. The fastest method to sample from the Dirichlet is to draw y_1, y_2, y_3 from independent gamma distribution with common scale and shape parameters $a_1 = a_0 * x_1, a_2 = a_0 * x_2, a_3 = a_0 * x_3$ where for each yi, $x_i' = yi / \sum_{i=1}^{3} y_i$ [131].

We propose values for the parameters from a Dirichlet with means equal to the current parameter values and a_0 is a scaling parameter that determines how large a change to the parameters we propose. Once the new set of degradation parameters is proposed, the Metropolis Hasting ratio is calculated as

$$R = \left\{ 1, \left[\prod_{i=1}^{n} \prod_{j=1}^{m} \frac{D_{ij}(x')}{D_{ij}(x)} \right] \times \frac{\frac{\Gamma(\sum_{i=1}^{3} a_0 x_i')}{\prod_{i=1}^{3} \Gamma(a_0 x_i')} \times \prod_{i=1}^{3} (x_i)^{a_0 x_i' - 1}}{\frac{\Gamma(\sum_{i=1}^{3} a_0 x_i)}{\prod_{i=1}^{3} \Gamma(a_0 x_i)} \times \prod_{i=1}^{3} (x_i')^{a_0 x_i - 1}} \right\}, \tag{4.4}$$

with the likelihood ratio evaluates across all sites m at hypothetical nodes n.

Next, a nucleotide at a particular site and hypothetical node is proposed to be changed. The likelihood ratio is the product of two fractions. First fraction is given by the substitution probabilities in degradation matrix $\mathbf{D}$ corresponding to the proposed and current nucleotide at the hypothetical node. The second fraction is the ratio of transition probabilities along the branch connecting the hypothetical node with its parent node, a process described by the uniformized substitution matrix $\mathbf{M}$. The acceptance ratio in this case is written as:

$$R = \left\{ 1, \frac{D_{a'b}}{D_{ab}} \times \frac{M_{ca'}}{M_{ca}}. \right\},$$

where a' and a are the proposed and current nucleotide at the randomly chosen hypothetical node, b is the nucleotide at the end of the edge connecting the hypothetical node to the fossil nucleotide and c is the nucleotide at the same site at the parent node of the chosen hypothetical node.

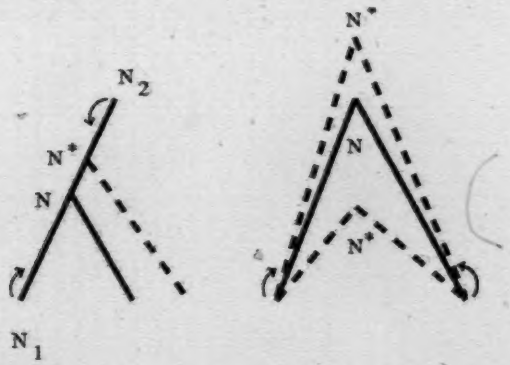


Figure 4.2: Algorithm for root node (left) or internal node sliding (right) for molecular clock implementation.

Simulating data

This novel model requires a different approach to simulate data for use in verifying its statistical performance. Because most extracted ancient DNA is mitochondrial with a more or less constant rate of evolution between and within species, we tested the model on clock trees only. The program I wrote for creating random trees can also generate random clock trees. Once a random tree is built, a random node is chosen, including the root node (the tree structure implemented in BYPASSR requires a rooted tree, but the root location is meaningless), and it is slid on the edge as depicted in Fig. 4.2, satisfying the condition that each node at the tip to be equidistant from the root. The green arrows indicate that the location of the sliding node is reflected back if the node tries to move below or above the adjacent node. The newick string with branch lengths included is imported in EVOLVER (PAML) [94] control file and a set of data with a specified α is generated. The next step is to assign a proportion of the sequences to be ancient. Another code I wrote continues to "evolve" the sequences for the nodes corresponding to ancient data. The parameters p, q and z are set to specific values and a proportion q of the total sites at the hypothetical nodes (randomly chosen) are allowed to "degrade". The location and the types of change are stored for post-analysis comparison.

4.3 Statistical performance

Designing the *in silico* experiments is challenging because of the poor prior information available from empirical studies and to formulate the aims of the study before generating data is essential. First, we want to find out how accurate the model is in recovering the sites we know are damaged. In addition, because we consider a mixture of extant sequences with some "degraded" ancient DNA, it is important to find the optimal proportion of fossil data necessary to recover the original nucleotides at the damaged sites

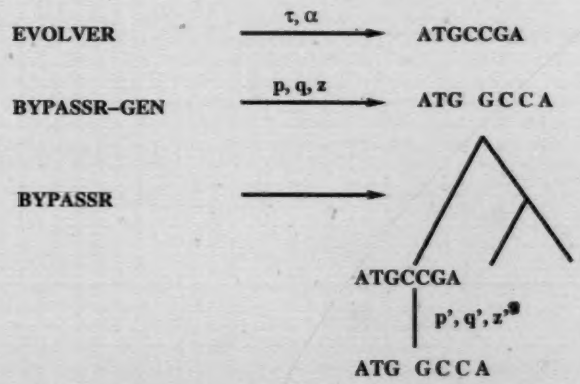


Figure 4.3: Generate data for degradation model testing.

(i.e. as existed before the degradation process is allowed to operate) and good estimates of the other parameters in the model. We also want to understand how site-specific rate estimates are affected by the presence of the damaged sites and the potential implications in detection of selection.

The ancient DNA sequences rarely exceed a length of a few hundred, so we generated data sets with 500 nucleotides only. The percent of degraded sites take values of 0.5%, 1% and 2%. We also vary the proportion of ancient DNA sequences from half to a fifth (s.a. 10 ancient sequences out of 20 sequences). For each of the dataset with the same number of sequences different random clock trees were used. We also let different degrees of rate variation among sites by setting α to 0.1, 0.5 and 1. When the "degradation" data sets are generated, the site and the fossil sequence where damage occurs are recorded. At the end of the analysis we have the list of sites that are different from the ones in the input sequences, but identical with initially generated sequences as shown in Fig. 4.3.

The input values of α in EVOLVER and their estimates in the "degraded" datasets are in columns 3 and 4 of Table 4.1. In all datasets α is overestimated, especially when α is smaller. Therefore, extreme rate heterogeneity among sites might be masked if there are 5% damaged sites. The same is true even of an α=0.5, but when there is less rate variation among sites the effect of damaged sites is not so visible. The proportion of sites not affected by degradation is greatly recovered in all the data set combinations, columns 4 and 6 Table 4.1. Columns 6 and 8 have the true and the estimated proportion of sites that underwent substitutions caused by degradation process and the column 9 gives the percent of how many sites were correctly found. In all the cases, more than 75% of the degraded sites were recovered. The majority of the posterior probabilities at the nucleotides at damaged sites are over 0.8. An interesting conclusion is that increasing the percent of degraded sites in the generated data improves the chance of BYPASSR finding them.

The simulation results show that our program is able to recover the sites that are

possibly modified by the degradation process. Being a stochastic process, the percent of correctly identified sites is not expected to be 100%. The best results are obtained when the number of degraded sites in the ancient sequences is large, when $q = 0.2$ meaning that 20% of the total fossil sites are modified by the degradation process. The percent of the sites we correctly identified to be degraded is between 90 and 95%. Reducing the number of degraded sites and implicitly the amount of information for the Markov process the percent drops to 80%.

We are interested in comparing the site-specific substitution rates inferred when damaged data is analyzed using a model that allows the presence of degraded sites with the estimates obtained using a model that does not account for the possibility of degraded sites. We expect a significant difference between the two. On the other hand, we expect to obtain similar results when we use degraded data with the degradation model implementation (BYPASSR-degr) and data without artifacts analyzed with the implementation of the model described in the previous chapters. The similarity is given by the correct assignment of the nucleotides at the degraded nodes and sites that recreates the sequences before the inclusion of the damaged nucleotides. As Fig. 4.4 shows, our expectations are met and we notice an important difference between the posterior mean rates when the data with incorporated errors is analyzed using a model that integrates over the uncertainty in the fossil data versus a model that ignores the damaged nucleotides (A panels). Fig. 4.4 panels B shows the correlation between the posterior mean substitution rates when damaged data is analyzed allowing the degradation model and site-specific substitution rates estimates obtained from clean data analyzed with the "regular" BYPASSR.

Increasing the number of sequences, but keeping the same number of sequences of ancient DNA give comparable proportion of recovered damaged sites, but improves the strength of correlation between the mean posterior rates obtained with damaged data + degradation model and clean data + "regular" BYPASSR.

Next, we tested how the method behaves when half of the sequences are ancient and half are extant sequences. We generated 4 datasets of 50, 60, 70 and 80 sequences with $p = 0.9$, $q = 0.1$ and $\alpha = 0.5$. We obtained back the nucleotides at the damaged sites in proportion of 97.5 to 99.2% (see Table 4.3). This observation is in agreement with our expectation. Increasing the number of ancient sequences add more information about the degradation model, therefore its parameters are better estimated. But, such a large number of ancient DNA sequences, increases uncertainty in the substitution rate estimates. Increasing the number of sequences, but keeping the proportion of ancient DNA to be 0.5, the strong positive correlation is maintained between the estimates of substitution rates when damaged data is analyzed using the correct model and the those obtained using clean data (without amplification errors) analyzed with BYPASSR without a degradation model. This observation is shown in Fig. 4.5 panels B. The impact on the site-specific rates estimates is obvious when the data set containing half of the sequences, ancient DNA is analyzed without considering the presence of damaged sites.

To conclude, we propose a novel approach to deal with the degradation process and incorporate this process into the context of the continuous variation of substitution rates

among sites. In the limited analysis on simulated data, BYPASSR-degr, the implementation of the model proposed by us, performed very well identifying the damaged sites even if there are many damaged sites spread across a number of sequences and obtaining good estimates of the other tree parameters (i.e. branch lengths, site-specific rate, GTR model parameters etc.). An efficient recovery of the tree parameters is possible when the number of fossil sequences is large enough that a model of degradation is well defined, but is small enough, when compared to the number of extant sequences, such that information about the underlying substitution process along the tree is not lost.

		fossils	α	α degr	true p	true q	est. p	est. q	found	not found	extra
set1	20seqs	10	0.1	0.307	0.95	0.04	0.957	0.035	0.844	0.157	0.025
	30seqs	10	0.1	0.241	0.95	0.04	0.960	0.035	0.879	0.121	0.046
	40seqs	10	0.1	0.235	0.95	0.04	0.964	0.029	0.826	0.174	0.037
	50seqs	10	0.1	0.204	0.95	0.04	0.949	0.043	0.921	0.079	0.049
set2	20seqs	10	0.1	0.304	0.80	0.15	0.812	0.141	0.958	0.042	0.041
	30seqs	10	0.1	0.249	0.80	0.15	0.803	0.153	0.940	0.060	0.047
	40seqs	10	0.1	0.316	0.80	0.15	0.774	0.161	0.909	0.091	0.174
	50seqs	10	0.1	0.176	0.80	0.15	0.800	0.152	0.911	0.089	0.051
set3	20seqs	10	0.5	0.533	0.95	0.04	0.948	0.041	0.770	0.230	0.164
	30seqs	10	0.5	0.565	0.95	0.04	0.955	0.035	0.790	0.210	0.163
	40seqs	10	0.5	0.554	0.95	0.04	0.947	0.042	0.858	0.142	0.133
	50seqs	10	0.5	0.523	0.95	0.04	0.944	0.048	0.773	0.227	0.145
set4	20seqs	10	0.5	0.738	0.80	0.15	0.811	0.140	0.944	0.056	0.093
	30seqs	10	0.5	0.498	0.80	0.15	0.797	0.142	0.942	0.058	0.093
	40seqs	10	0.5	0.564	0.80	0.15	0.798	0.155	0.933	0.067	0.067
	50seqs	10	0.5	0.583	0.80	0.15	0.794	0.155	0.953	0.047	0.079
set5	20seqs	10	1	1.014	0.95	0.04	0.927	0.042	0.918	0.082	0.358
	30seqs	10	1	1.066	0.95	0.04	0.957	0.027	0.823	0.177	0.227
	40seqs	10	1	1.061	0.95	0.04	0.958	0.035	0.818	0.182	0.143
	50seqs	10	1	1.022	0.95	0.04	0.949	0.037	0.733	0.267	0.210
set6	20seqs	10	1	1.070	0.80	0.15	0.785	0.162	0.943	0.057	0.110
	30seqs	10	1	1.306	0.80	0.15	0.798	0.146	0.945	0.055	0.111
	40seqs	10	1	1.013	0.80	0.15	0.794	0.150	0.964	0.036	0.071
	50seqs	10	1	1.105	0.80	0.15	0.817	0.144	0.923	0.077	0.066

Table 4.1: Estimates for the parameters of the degradation model.

	fossils	α	α degr	true p	true q	est. p	est. q	found	not found	extra
50seqs	25	0.5	0.464	0.90	0.10	0.901	0.097	0.975	0.0250	0.030
60seqs	30	0.5	0.473	0.90	0.10	0.889	0.111	0.980	0.020	0.022
70seqs	35	0.5	0.470	0.90	0.10	0.902	0.096	0.992	0.008	0.025
80seqs	40	0.5	0.408	0.90	0.10	0.902	0.097	0.987	0.013	0.015

Table 4.2: Estimates for the parameters of the degradation model when half of the sequences are ancient DNA.

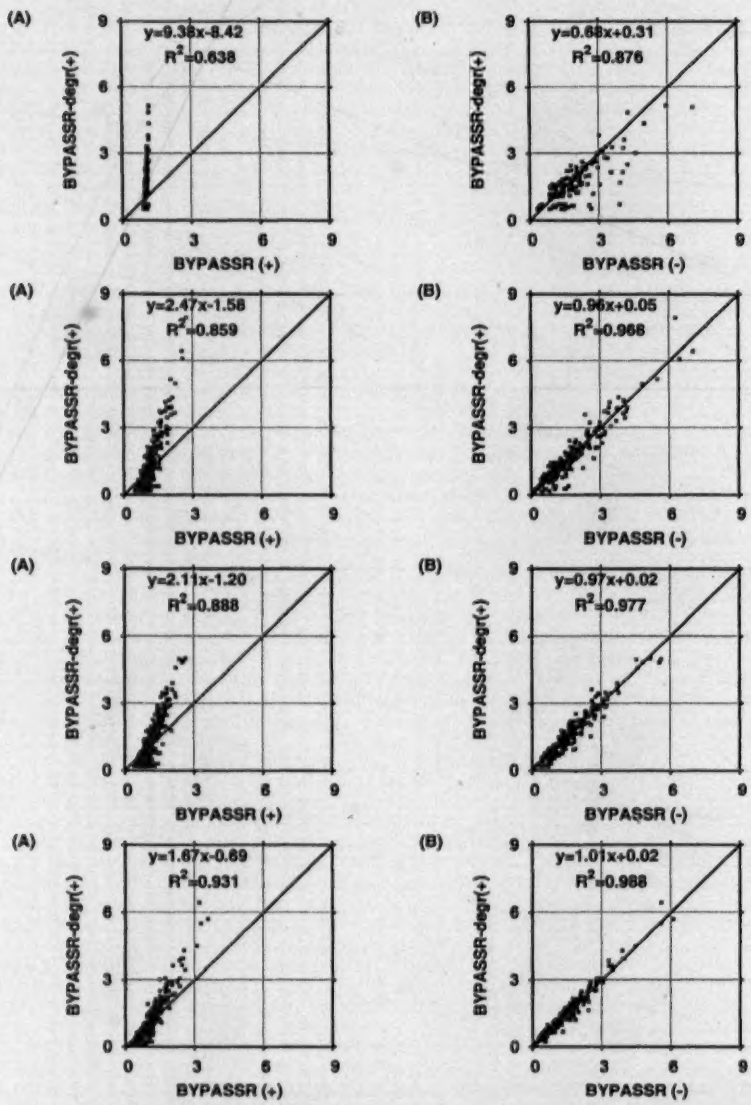


Figure 4.4: Comparison between the mean posterior substitution rates when degraded data (+) is analyzed with BYPASSR-degr and BYPASSR (A panels). B panels show the correlation between the mean posterior rates obtained from degraded data using BYPASSR-degr versus original data set before including the errors (-) analyzed with BYPASSR. The simulated data sets have 20, 30, 40 and 50 sequences (from top to bottom). (set 4 in Table 4.1).

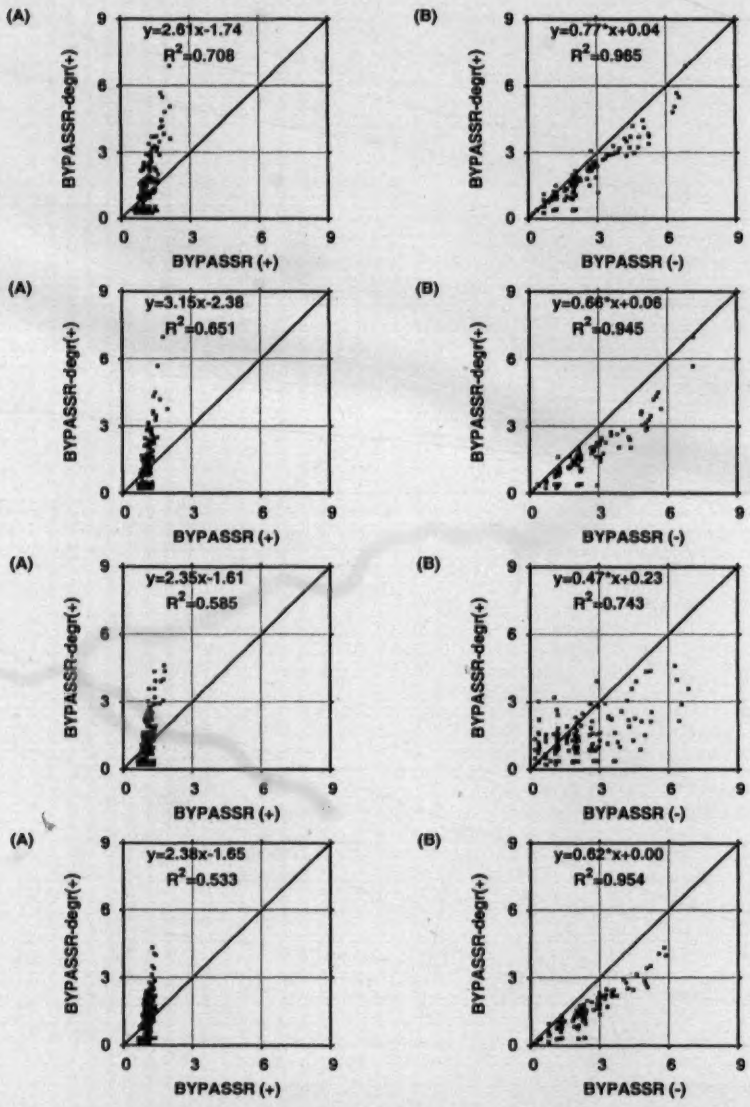


Figure 4.5: Comparison between the mean posterior substitution rates when "degraded" data (+) is analyzed with BYPASSR-degr and BYPASSR (A panels). B panels show the correlation between the mean posterior rates obtained from degraded data using BYPASSR-degr versus original data set before including the errors (-) analyzed with BYPASSR. The simulated data sets have 50, 60, 70 and 80 sequences (from top to bottom). The proportion fossil/extant sequences is equal (Table 4.3).

Chapter 5

Conclusions

I have used a new technique for calculating substitution probabilities using complex models by uniformization of the Markov substitution process. An advantage of this formulation of the transition probabilities is that it allows efficient augmentation of the data in a MCMC analysis by treating the substitution events as random variables in the chain and eliminating the need to numerically calculate the transition probabilities in complex substitution models by use of matrix exponentiation. The method is applied to infer site-specific rates and a program, BYPASSR, is presented. The method provides estimates of branch lengths that agree closely with those inferred by empirical Bayes methods using a discrete gamma approximation implemented in the program BASEML. However, the discrete gamma approximation appears to cause systematic underestimates of rates for rapidly evolving sites unless a large number of rate categories are used. My analyses of the posterior distributions of site-specific rates suggest that a large number of taxa are needed to accurately infer rates. These findings agree with previous analyzes of the effect of taxon sampling on estimates of site-specific rates using simplified models by [132]. As the number of rate categories in the discrete gamma approximation is increased, the site-specific rate estimates obtained using BASEML approach more closely those obtained using BYPASSR. In the cases with extreme values, the most interesting one for the study of selection, the maximum number of categories allowed by BASEML does not suffice the fit the wide range of rate values. The tendency of overestimating the before last category and underestimating the last category is perpetuated. Although, more studies has to be done on the subject, my observations may be extended to other methods that involve discretization of substitution rates. More specifically, the dN/dS ratio is described as a continuous parameter in all the available parametric methods, but in fact, independent of the choice of distribution, an approximation is used instead. A continuous distribution is discretized into a limited number of categories.

The abundance of sophisticated methods to analyze protein-coding DNA has identified many genes or codons within genes to be marked by positive selection. However, the availability of comprehensive methods for the majority of the DNA, which is noncoding, is extremely limited. My method is not restricted to the analysis of protein coding DNA. I demonstrated that the genetic information available in the noncoding

DNA clearly distinguished between a slowly and a fast evolving site. The possibility of a fast evolving site to be a positively selected is marked by the mutational load of the noncoding sites. However, there is little doubt about the slowly evolving sites. Regulatory regions with remarkable degree of conservation among species, usually found only through the tools of comparative genomics, have proved essential in understanding the functions of many genes.

I applied my method to datasets previously analyzed for the possibility of containing positively selected sites. HIV-1 *pol* polypeptide and class I major human histocompatibility genes are known for having many such sites. The results obtained by us identified the nucleotide sites belonging to these codons, as being good candidate for positive selection. Even more, I consider that the few additional sites found be us to also be potential positively selected, escaped detection by the methods that considered the dN/dS rate ratio as the only mark of selection. The use of a true continuous distribution to describe the substitution rates reduced the number of approximations made by the probabilistic models. In this sense, my model may be more biologically realistic than a codon based model that force the substitution rates to take limited number of values.

The parametric methods for phylogenetic inference have a limited number of sequences that can be analyzed at once. The limitation is given by the numerous calculations required to reevaluate the whole system after a small change to one of the thousands (or more) parameters in it. The uniformization technique greatly reduces the magnitude of these calculations and it allowed us to analyze a large data set: 688 sequences of mammalian cytochrome *b* and identify sites that might evolve under one of modeling forces of nature, positive selection.

I also proposed a model to account for the uncertainty in the ancient DNA data. The capacity of recovering ancient DNA improved greatly in the recent years, but a few amplification errors can appear during the processing of ancient DNA. My model addresses to these kinds of errors. By allowing data to decide if the observed nucleotides at the ancient DNA sequences have better probability than any of the other three nucleotides, conditional on a discrete Markov process suggested by us to mimic the degradation process, I provided a measure of the quality of phylogenetic inference based on ancient DNA.

Moreover, the general approach of uniformization should have broad application in phylogenetic inference, potentially allowing much more complex substitution models to be efficiently implemented. I have demonstrated the usefulness of uniformization and data augmentation for the specific problem of modeling among-site rate variation. However, the method should be useful for modeling any substitution process for which a continuous-time rate matrix can be specified. This might include complicated models of dependence between sites, etc. One obvious extension that will be very efficient would be to simultaneously model both among-site rate variation and among-lineage rate variation. The same advantages incurred when modeling among-site rate variation will apply here also (e.g., no need to recalculate the discrete matrix product when a change is proposed for a lineage specific rate, etc).

Bibliography

[1] H. Badrane and N. Tordo. Host switching in Lyssavirus history from the Chiroptera to the Carnivora orders. *J Virol*, 75(17):8096–104, 2001.

[2] J. Zhang and H. F. Rosenberg. Complementary advantageous substitutions in the evolution of an antiviral RNase of higher primates. *Proc Natl Acad Sci U S A*, 99(8):5486–91, 2002.

[3] A. P. Martin and S. R. Palumbi. Body size, metabolic rate, generation time, and the molecular clock. *Proc Natl Acad Sci U S A*, 90(9):4087–91, 1993.

[4] P. M. Sharp and G. Matassi. Codon usage and genome evolution. *Curr Opin Genet Dev*, 4(6):851–60, 1994.

[5] E. Zuckerkandl and L. Pauling. *Horizons in Biochemistry, Molecular disease, evolution, and genetic heterogeneity*. Academic Press, New York, 1962.

[6] J. Huelsenbeck and B. Rannala. Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. *Syst Biol*, 53(6):904–13, 2004.

[7] A. R. Lemmon and E. C. Moriarty. The importance of proper model assumption in bayesian phylogenetics. *Syst Biol*, 53(2):265–77, 2004.

[8] Z. Yang. Among-site rate variation and its impact on phylogenetic analysis. *Trends in Ecology and Evolution*, (11):367–372, 1996.

[9] Z. Yang and S. Kumar. Approximate methods for estimating the pattern of nucleotide substitution and the variation of substitution rates among sites. *Mol Biol Evol*, 13(5):650–9, 1996.

[10] J. Sullivan, K. E. Holsinger, and C. Simon. The effect of topology on estimates of among-site rate variation. *J Mol Evol*, 42(2):308–12, 1996.

[11] Z. Yang. *Handbook of statistical genetics, chap. 9. Adaptive molecular evolution, pages 229–254*. 2003.

[12] N. Goldman. Statistical tests of models of DNA substitution. *J Mol Evol*, 36(2):182–98, 1993.

[13] J. P. Huelsenbeck and B. Rannala. Phylogenetic methods come of age: testing hypotheses in an evolutionary context. *Science*, 276(5310):227–32, 1997.

[14] J. Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol*, 17(6):368–76, 1981.

[15] Z. Yang. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol Biol Evol*, 10(6):1396–401, 1993.

[16] F. Rodriguez, J. L. Oliver, A. Marin, and J. R. Medina. The general stochastic model of nucleotide substitution. *J Theor Biol*, 142(4):485–501, 1990.

[17] T. H. Jukes and C. R. Cantor. Evolution of protein molecules. In HN Munro, editor, *Mammalian protein metabolism*, volume II, pages 21–132. Academic Press, New York, 1969.

[18] M. Kimura. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol*, 16(2):111–20, 1980.

[19] D. L. Swofford, G.J. Olsen, P.J. Waddell, and D.M. Hillis. *Molecular Systematics, Second Ed. Phylogenetic inference*. Sinauer Associates, Sunderland, Mass., 1996.

[20] M. O. Dayhoff, R. M. Schwartz, and B. C. Orcutt. *Atlas of protein sequence and structure*, volume 5,Suppl. 3. National Biomedical Research Foundation, Washington, D.C, 1978.

[21] S. Henikoff and J. G. Henikoff. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*, 89(22):10915–9, 1992.

[22] D. T. Jones, W. R. Taylor, and J. M. Thornton. The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci*, 8(3):275–82, 1992.

[23] S. Whelan and N. Goldman. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol Biol Evol*, 18(5):691–9, 2001.

[24] M. Hasegawa and T. Hashimoto. Ribosomal RNA trees misleading? *Nature*, 361(6407):23, 1993.

[25] Z. Yang and D. Roberts. On the use of nucleic acid sequences to infer early branchings in the tree of life. *Mol Biol Evol*, 12(3):451–8, 1995.

[26] S. V. Muse and B. S. Gaut. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol Biol Evol*, 11(5):715–24, 1994.

[27] N. Goldman and Z. Yang. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol*, 11(5):725–36, 1994.

[28] Z. Yang, R. Nielsen, and M. Hasegawa. Models of amino acid substitution and applications to mitochondrial protein evolution. *Mol Biol Evol*, 15(12):1600–11, 1998.

[29] Z. Yang, R. Nielsen, N. Goldman, and A. M. Pedersen. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, 155(1):431–49, 2000.

[30] Z. Yang and R. Nielsen. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *J Mol Evol*, 46(4):409–18, 1998.

[31] S. K. Pond and S. V. Muse. Site-to-site variation of synonymous substitution rates. *Mol Biol Evol*, 22(12):2375–85, 2005.

[32] Z. Yang. Maximum-likelihood models for combined analyses of multiple sequence data. *J Mol Evol*, 42(5):587–96, 1996.

[33] B. Shapiro, A. Rambaut, and A. J. Drummond. Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. *Mol Biol Evol*, 23(1):7–9, 2006.

[34] M. Z. Ludwig. Functional evolution of noncoding DNA. *Curr Opin Genet Dev*, 12(6):634–9, 2002.

[35] C. I. Castillo-Davis. The evolution of noncoding DNA: how much junk, how much func? *Trends Genet*, 21(10):533–6, 2005.

[36] G. V. Kryukov, S. Schmidt, and S. Sunyaev. Small fitness effect of mutations in highly conserved non-coding regions. *Hum Mol Genet*, 14(15):2221–9, 2005.

[37] W. Huang, D. M. Umbach, and L. Li. Accurate anchoring alignment of divergent sequences. *Bioinformatics*, 22(1):29–34, 2006.

[38] J. Wang, P. D. Keightley, and T. Johnson. MCALIGN2: Faster, accurate global pairwise alignment of non-coding DNA sequences based on explicit models of indel evolution. *BMC Bioinformatics*, 7(1):292, 2006.

[39] W. S. Wong and R. Nielsen. Detecting selection in noncoding regions of nucleotide sequences. *Genetics*, 167(2):949–58, 2004.

[40] M. Hasegawa, H. Kishino, and T. Yano. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol*, 22(2):160–74, 1985.

[41] W. M. Fitch and E. Markowitz. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. *Biochem Genet*, 4(5):579–93, 1970.

[42] T. Uzzell and K. W. Corbin. Fitting discrete probability distributions to evolutionary events. *Science*, 172(988):1089–96, 1971.

[43] M. Nei, R. Chakraborty, and P. A. Fuerst. Infinite allele model with varying mutation rate. *Proc Natl Acad Sci U S A*, 73(11):4164–8, 1976.

[44] G. B. Golding. Estimates of DNA and protein sequence divergence: an examination of some assumptions. *Mol Biol Evol*, 1(1):125–42, 1983.

[45] M. Nei and T. Gojobori. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol*, 3(5):418–26, 1986.

[46] L. Jin and M. Nei. Limitations of the evolutionary parsimony method of phylogenetic analysis. *Mol Biol Evol*, 7(1):82–102, 1990.

[47] G. J. Olsen. Earliest phylogenetic branchings: comparing rRNA-based evolutionary trees inferred with various techniques. *Cold Spring Harb Symp Quant Biol*, 52:825–37, 1987.

[48] Z. Yang. *Computational Molecular Evolution*. Oxford University Press, 2006.

[49] J. Felsenstein. Taking variation of evolutionary rates between sites into account in inferring phylogenies. *J Mol Evol*, 53(4-5):447–55, 2001.

[50] Z. Yang. Maximum likelihood phylogenetic estimation from dna sequences with variable rates over sites: approximate methods. *J Mol Evol*, 39(3):306–14, 1994.

[51] J. H. Gillespie. *The causes of molecular evolution*. Oxford University Press New York, 1991.

[52] W-H. Li. *Molecular evolution*. Sinauer Sunderland, MA, 1997.

[53] W. M. Fitch. Toward defining the course of evolution: Minimum change for a specified tree topology. *Syst Zool*, 20:406–416, 1971.

[54] J. Zhang, R. Nielsen, and Z. Yang. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol Biol Evol*, 22(12):2472–9, 2005.

[55] R. Nielsen. Mapping mutations on phylogenies. *Syst Biol*, 51(5):729–39, 2002.

[56] A. Jensen. Markoff chains as an aid in the study of Markoff processes. *Skandinavisk Aktuarietidskrift*, (36):87–91, 1953.

[57] S.M. Ross. *Stochastic processes*. Willey, New York, 1983.

[58] J. Felsenstein. *Inferring Phylogenies*. Sinauer, Sunderland, 2004.

[59] D. M. Robinson, D. T. Jones, H. Kishino, N. Goldman, and J. L. Thorne. Protein evolution with dependence among codons due to tertiary structure. *Mol Biol Evol*, 20(10):1692–704, 2003.

[60] B. Gaschen, J. Taylor, K. Yusim, B. Foley, F. Gao, D. Lang, V. Novitsky, B. Haynes, B. H. Hahn, T. Bhattacharya, and B. Korber. Diversity considerations in HIV-1 vaccine selection. *Science*, 296(5577):2354–60, 2002.

[61] S. A. R. S. Chinese, Molecular Epidemiology Consortium. Molecular evolution of the SARS coronavirus during the course of the SARS epidemic in China. *Science*, 303(5664):1666–9, 2004.

[62] J. W. Thornton, E. Need, and D. Crews. Resurrecting the ancestral steroid receptor: ancient origin of estrogen signaling. *Science*, 301(5640):1714–7, 2003.

[63] T. Gabaldon and M. A. Huynen. Reconstruction of the proto-mitochondrial metabolism. *Science*, 301(5633):609, 2003.

[64] C. B. Stewart, J. W. Schilling, and A. C. Wilson. Adaptive evolution in the stomach lysozymes of foregut fermenters. *Nature*, 330(6146):401–4, 1987.

[65] B. S. Chang, K. Jonsson, M. A. Kazmi, M. J. Donoghue, and T. P. Sakmar. Recreating a functional ancestral archosaur visual pigment. *Mol Biol Evol*, 19(9):1483–9, 2002.

[66] D. Schluter. Uncertainty in ancient phylogenies. *Nature*, 377(6545):108–10, 1995.

[67] Z. Yang, S. Kumar, and M. Nei. A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics*, 141(4):1641–50, 1995.

[68] T. R. Scultz and G.A. Churchill. The role of subjectivity in reconstructing ancestral character states: A Bayesian approach to unknown rates, states, and transformation asymmetries. *Syst. Biol.*, 48(651-664), 1999.

[69] J. P. Huelsenbeck and J. P. Bollback. Empirical and hierarchical Bayesian estimation of ancestral states. *Syst Biol*, 50(3):351–66, 2001.

[70] Z. Yang. Paml: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci*, 13(5):555–6, 1997.

[71] J. Zhang and M. Nei. Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods. *J Mol Evol*, 44 Suppl 1:S139–46, 1997.

[72] N. M. Krishnan, H. Seligmann, C. B. Stewart, A. P. De Koning, and D. D. Pollock. Ancestral sequence reconstruction in primate mitochondrial DNA: compositional bias and effect on functional inference. *Mol Biol Evol*, 21(10):1871–83, 2004.

[73] W. Cai, J. Pei, and N. V. Grishin. Reconstruction of ancestral protein sequences and its applications. *BMC Evol Biol*, 4:33, 2004.

[74] J. W. Thornton. Resurrecting ancient genes: experimental analysis of extinct molecules. *Nat Rev Genet*, 5(5):366–75, 2004.

[75] P. D. Williams, D. D. Pollock, B. P. Blackburne, and R. A. Goldstein. Assessing the accuracy of ancestral protein reconstruction methods. *PLoS Comput Biol*, 2(6):e69, 2006.

[76] R. Higuchi, B. Bowman, M. Freiberger, O. A. Ryder, and A. C. Wilson. DNA sequences from the quagga, an extinct member of the horse family. *Nature*, 312(5991):282–4, 1984.

[77] S. Paabo. Molecular cloning of Ancient Egyptian mummy DNA. *Nature*, 314(6012):644–5, 1985.

[78] S. Paabo, H. Poinar, D. Serre, V. Jaenicke-Despres, J. Hebler, N. Rohland, M. Kuch, J. Krause, L. Vigilant, and M. Hofreiter. Genetic analyses from ancient DNA. *Annu Rev Genet*, 38:645–79, 2004.

[79] S. Paabo. Ancient DNA: extraction, characterization, molecular cloning, and enzymatic amplification. *Proc Natl Acad Sci U S A*, 86(6):1939–43, 1989.

[80] A. Hansen, E. Willerslev, C. Wiuf, T. Mourier, and P. Arctander. Statistical evidence for miscoding lesions in ancient DNA templates. *Mol Biol Evol*, 18(2):262–5, 2001.

[81] J. Binladen, C. Wiuf, M. T. Gilbert, M. Bunce, R. Barnett, G. Larson, A. D. Greenwood, J. Haile, S. Y. Ho, A. J. Hansen, and E. Willerslev. Assessing the fidelity of ancient DNA sequences amplified from nuclear genes. *Genetics*, 172(2):733–41, 2006.

[82] M. Hofreiter, V. Jaenicke, D. Serre, A. Haeseler Av, and S. Paabo. DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA. *Nucleic Acids Res*, 29(23):4793–9, 2001.

[83] M. T. Gilbert, E. Willerslev, A. J. Hansen, I. Barnes, L. Rudbeck, N. Lynnerup, and A. Cooper. Distribution patterns of postmortem damage in human mitochondrial DNA. *Am J Hum Genet*, 72(1):32–47, 2003.

[84] M.T.P. Gilbert, B. Shapiro, A. Drummond, and A. Cooper. Post-mortem DNA damage hotspots in Bison (*Bison bison*) provide evidence for both damage and mutational hotspots in human mitochondrial DNA. *Journal of Archaeological Science*, 32:1053–1060, 2005.

[85] M. Banerjee and T.A. Brown. Non-random DNA damage resulting from heat treatment: implications for sequence analysis of ancient DNA. *Journal of Archaeological Science*, 31:59–63, 2004.

[86] J. L. Jensen and A. M. K Pedersen. Probabilistic models of DNA sequence evolution with context dependent rates of substitution. *Advances in Applied Probability*, (32):499–517, 2000.

[87] A. M. Pedersen and J. L. Jensen. A dependent-rates model and an MCMC-based methodology for the maximum-likelihood analysis of sequences with overlapping reading frames. *Mol Biol Evol*, 18(5):763–76, 2001.

[88] D. G. Hwang and P. Green. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc Natl Acad Sci U S A*, 101(39):13994–4001, 2004.

[89] L. Mateiu and B. Rannala. Inferring complex DNA substitution processes on phylogenies using uniformization and data augmentation. *Syst Biol*, 55(2):259–69, 2006.

[90] D. Zwickl and M. Holder. Model parameterization, prior distributions, and the general time-reversible model in Bayesian phylogenetics. *Syst Biol*, 53(6):877–88, 2004.

[91] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *J Chem Phys*, 21(6):1087–1092, 1953.

[92] K. Tamura and M. Nei. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol*, 10(3):512–26, 1993.

[93] J. Felsenstein. Distance methods for inferring phylogenies: a justification. *Evolution*, 38:16–24, 1984.

[94] Z. Yang. Phylogenetic analysis by maximum likelihood (PAML). version 3.13., 2002.

[95] R.A. Horn and C.R. Johnson. *Matrix Analysis*. Cambridge University Press, 1990.

[96] Z. Yang and B. Rannala. Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. *Mol Biol Evol*, 23(1):212–26, 2006.

[97] Z. Yang and J. P. Bielawski. Statistical methods for detecting molecular adaptation. *Trends in Ecol and Evolution*, 15(12):496–503, 2000.

[98] Z. Yang and T. Wang. Mixed model analysis of DNA sequence evolution. *Biometrics*, 51(2):552–61, 1995.

[99] M. Hasegawa, H. Kishino, and T. Yano. Man's place in hominoidea as inferred from molecular clocks of dna. *J Mol Evol*, 26(1-2):132–47, 1987.

[100] W. Yang, J. P. Bielawski, and Z. Yang. Widespread adaptive evolution in the human immunodeficiency virus type 1 genome. *J Mol Evol*, 57(2):212–21, 2003.

[101] http://evolution.genetics.washington.edu/phylip/newicktree.html.

[102] http://www.bioperl.org.

[103] http://www.gnuplot.info/.

[104] F. Ronquist and J. P. Huelsenbeck. Mrbayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19(12):1572–1574, 2003.

[105] S. A. Seibert, C. Y. Howell, M. K. Hughes, and A. L. Hughes. Natural selection on the gag, pol, and env genes of human immunodeficiency virus 1 (HIV-1). *Mol Biol Evol*, 12(5):803–13, 1995.

[106] R. Nielsen and Z. Yang. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics*, 148(3):929–36, 1998.

[107] http://abacus.gene.ucl.ac.uk/ziheng/data/YNGP2000.tar.gz.

[108] Z. Yang, W. S. Wong, and R. Nielsen. Bayes empirical bayes inference of amino acid sites under positive selection. *Mol Biol Evol*, 22(4):1107–18, 2005.

[109] T. Massingham and N. Goldman. Detecting amino acid sites under positive selection and purifying selection. *Genetics*, 169(3):1753–62, 2005.

[110] J. P. Huelsenbeck, S. Jain, S. W. Frost, and S. L. Pond. A Dirichlet process model for detecting positive selection in protein-coding DNA sequences. *Proc Natl Acad Sci U S A*, 103(16):6263–8, 2006.

[111] A. L. Hughes and M. Nei. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature*, 335(6186):167–70, 1988.

[112] Z. Yang and W. J. Swanson. Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. *Mol Biol Evol*, 19(1):49–57, 2002.

[113] R. Sainudiin, W. S. Wong, K. Yogeeswaran, J. B. Nasrallah, Z. Yang, and R. Nielsen. Detecting site-specific physicochemical selective pressures: applications to the Class I HLA of the human major histocompatibility complex and the SRK of the plant sporophytic self-incompatibility system. *J Mol Evol*, 60(3):315–26, 2005.

[114] W. S. Wong, R. Sainudiin, and R. Nielsen. Identification of physicochemical selective pressure on protein encoding nucleotide sequences. *BMC Bioinformatics*, 7:148, 2006.

[115] http://abacus.gene.ucl.ac.uk/ziheng/data.html.

[116] M. Poidinger, R. A. Hall, and J. S. Mackenzie. Molecular characterization of the Japanese encephalitis serocomplex of the flavivirus genus. *Virology*, 218(2):417–21, 1996.

[117] T. J. Chambers, C. S. Hahn, R. Galler, and C. M. Rice. Flavivirus genome organization, expression, and replication. *Annu Rev Microbiol*, 44:649–88, 1990.

[118] C. Woodgett and J. K. Rose. Amino-terminal mutation of the vesicular stomatitis virus glycoprotein does not affect its fusion activity. *J Virol*, 59(2):486–9, 1986.

[119] J. M. Coll. The glycoprotein G of rhabdoviruses. *Arch Virol*, 140(5):827–51, 1995.

[120] T. Mebatsion, M. J. Schnell, J. H. Cox, S. Finke, and K. K. Conzelmann. Highly stable expression of a foreign gene from rabies virus vectors. *Proc Natl Acad Sci U S A*, 93(14):7310–4, 1996.

[121] E. V. Ravkov, J. S. Smith, and S. T. Nichol. Rabies virus glycoprotein gene contains a long 3' noncoding region which lacks pseudogene properties. *Virology*, 206(1):718–23, 1995.

[122] D. A. McClellan and K. G. McCracken. Estimating the influence of selection on the variable amino acid sites of the cytochrome B protein functional domains. *Mol Biol Evol*, 18(6):917–25, 2001.

[123] B. Prusak and T. Grzybowski. Non-random base composition in codons of mitochondrial cytochrome *b* gene in vertebrates. *Acta Biochim Pol*, 51(4):897–905, 2004.

[124] J. Castresana. Cytochrome b phylogeny and the taxonomy of great apes and mammals. *Mol Biol Evol*, 18(4):465–71, 2001.

[125] H. Zhang, D. Bhattacharya, and S. Lin. Phylogeny of dinoflagellates based on mitochondrial cytochrome b and nuclear small subunit rDNA sequence Comparisons. *J. Phycol.*, 41:411–420, 2005.

[126] Z. Yang. Relating physicochemical properties of amino acids to variable nucleotide substitution patterns among sites. *Pac Symp Biocomput*, pages 81–92, 2000.

[127] D. A. McClellan, E. J. Palfreyman, M. J. Smith, J. L. Moss, R. G. Christensen, and J. K. Sailsbery. Physicochemical evolution and molecular adaptation of the cetacean and artiodactyl cytochrome *b* proteins. *Mol Biol Evol*, 22(3):437–55, 2005.

[128] J. D. Thompson, D. G. Higgins, and T. J. Gibson. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, 22(22):4673–80, 1994.

[129] P. Rice, I. Longden, and A. Bleasby. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet*, 16(6):276–7, 2000.

[130] E. Willerslev, A. J. Hansen, J. Binladen, T. B. Brand, M. T. Gilbert, B. Shapiro, M. Bunce, C. Wiuf, D. A. Gilichinsky, and A. Cooper. Diverse plant and animal genetic records from Holocene and Pleistocene sediments. *Science*, 300(5620):791–5, 2003.

[131] A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin. *Bayesian Data Analysis*. Chapman Hall/CRC, 2004.

[132] D. D. Pollock and W. J. Bruno. Assessing an unknown evolutionary process: effect of increasing site-specific knowledge through taxon addition. *Mol Biol Evol*, 17(12):1854–8, 2000.

[133] G. Perrière and M. Gouy. WWW-Query: An on-line retrieval system for biological sequence banks. *Biochimie*, 78:364–469, 1996.

[134] J. Sullivan and P. Joyce. Model selection in phylogenetics. *Ann Rev Ecol Evol Syst*, pages 36:445–466, 2005.

Appendix A

BYPASSR v.1.0

A.1 Overview

BYPASSR (BaYesian Phylogenetic Analysis of Site-Specific Rates) is the implementation of a new method for inferring branch lengths and site-specific substitution rates from nucleotide sequences. It makes use of the Bayesian Markov Chain Monte Carlo method and of the uniformization (randomization) technique to calculate sequence substitution probabilities in complicated DNA substitution models such as general time reversible model (GTR). The new feature in BYPASSR is the allowance of continuous gamma distribution to model site-specific rates even for large phylogenetic trees. The detailed model description is available at *Syst. Biol. 55(2):259-269,2006. L. Mateiu and B. Rannala. Inferring Complex DNA Substitution Processes on Phylogenies Using Uniformization and Data Augmentation.*

A.2 Installation notes

BYPASSR is a command-line driven computer program written in C/C++.

BYPASSR is available for Unix-based operating systems (Linux, MacOS X). For Windows users, BYPASSR was tested on **CYGWIN**. As an alternative for Windows users, LINUX can be installed as a guest operating system by using **VMware**. VMware player is freely distributed. BYPASSR automatically creates plots (as pdf files) if **GNUPLOT** is installed. **GNUPLOT** is a free, very flexible plotting software. There is also available a parallel version on BYPASSR that requires **LAM/MPI** installlation. LAM makes no restriction on what unix-based machines you can run on. The parallel version of BYPASSR can be used on Linux cluster, a dual processor machine or even on single processor, regular PCs. There is a number of advantages (detailed in the Section 2.3.8) by running multiple independent MCMC chains and I recommend the parallel version.

After unpacking, create the executable

```
bypassr
```

with

```
[...]$ make
```

The compilation requires a gcc compiler, version greater than 3.2. During compilation, the program searches for the lam and Gnuplot installation and it will print on screen if any of them is available. If **lam/mpi** is installed on the machine and it is on user PATH, the executable for parallel use is also created **bypassr-mpi**). Next, you can copy **bypassr** or **bypassr-mpi** to your working directory or add it the user PATH for system to find it. Also, you need to copy the file called **controlfile** in your working directory. Run the program by typing in a Unix shell

```
[...]$ ./bypassr
```

or for parallel version to run it on 3 cpus for example

```
[...]$ mpirun -np 3 bypassr-mpi
```

On some MacOS, after starting the executable **bypassr**, you might get an error like this SEGMENTATION FAULT. This requires a modification to the **Makefile**. Edit the **Makefile** by deleting the flag **-O3** following CC = g++ and do again **make**. If you get any other error during early stage of your analysis an *Errors.txt* file is created to suggest solutions to the problem or contact me at lmateiu@ualberta.ca.

A.3 Input Files

A.3.1 Data file

The input file must contain **aligned** nucleotides. It recognizes **T, t, C, c, A, a, G, g** and **-**. Any other character is treated as **-**. For now, the program is using clean data and will eliminate all the sites with at least one missing or ambiguous character. The program reads only FASTA format, from which you eliminated the comments and other sequence descriptions, but sequence name. It does not require a specific number of nucleotides on each row. Your file can be converted to FASTA by using **ReadSeq**, http://www.bioinformatics.vg/sms/**Sequence Manipulation Suite** or any other file format converter.

Example of a datafile used by BYPASSR:

```
>SeqName1
GCGGGGAACAGAGATGGAG?????GAACGCGGGAGCGGCCGTGAGGAAAGAGTG
AACAATAGGCCGGATCGGAACAAGGAGCG-----CACGCCTATAGCCAATAGGG
CAGGGCTCGCGAGCGCAAGCGAATGAGACGGGGAGACA
>SeqName2
GCGGGCAACAGAGATCGAGACTTGGAACGCGGGAGCGGCCGTGAGGAAGGAG--
AACAAGAGGCCGGATCGGAACAAGGAGGGGCACGCCTAGAGCCAGTAGGGGGGG
CACAGCTCGCGAGCGCCAGCGAATGCGACGCGGAGACA
```

We developed a model to account for the uncertainty in the ancient DNA caused by amplification errrors during PCR processing. The use of this model, requires the input of a mixture of extant DNA and ancient DNA. The ancient DNA need to have an age in years specified after the sequence name. The actual age is not used in the analysis as our model is a time independent process.

```
>AncientDNA 1000
GCGGGGAACAGAGA...
>Ancient DNA 20000
GCGGGCAACAGAG....
>SeqName1
TCCTCCATTACTT....
```

A.3.2 Tree file

BYPASSR does not do tree search, therefore the user has to provide a reasonable phylogenetic tree. There is a long list of the softwares that perform this task **here**. Some of most popular ones are PAUP, PHYLIP, MrBayes, PAML (for small trees) etc. The phylogenetic tree must be in **newick tree format** with branch lengths included. The program requires a rooted tree structure, but the root location is meaningless. If you are not sure if your tree is rooted or not, **NJplot** [133] can be used to save your tree as rooted.

Example of a tree file (the whole array must be in one line):

```
((SeqName1:0.01,(SeqName9:0.12,SeqName10:0.05):0.10):0.08,
((((SeqName5:0.06,SeqName6:0.00):0.02,SeqName4:0.02):0.03,
SeqName3:0.01):0.00,(SeqName2:0.00,(SeqName7:0.05,SeqName8
:0.01):0.03):0.02):0.03);
```

Note: The sequence names in the tree file must match exactly the sequence name in the data file. Special characters such as **( ) : ; ,** are not allowed in the sequence name.

A.4 Control file controlfile

A.4.1 BYPASSR

This file specifies the input files and a bunch of user-defined variables.

- **<input datafile>** DNA sequence file in fasta format. If the file is not in the same directory with *bypassr*, write the complete path for the datafile.
- **<input treefile>** Input the file name containing one binary rooted phylogenetic tree in newick format.

- **<seed>** A large integer used by the random number generator. The options are:

  RANDOM The program will take a large integer based on the current time of the machine.

  188732876 Any integer from 1 to 2147483647.

- **<DNA substitution model>** Choose the preferred combination of parameters for the DNA model of sequence evolution. The transition/transvertion rate ratio is obtained in the same way as in PAML [94]

GTR [16]	The most general model with variable base frequencies ($\pi_T \neq \pi_C \neq \pi_A \neq \pi_G$) and 6 relative rates (T⇄C) = a, (T⇄A) = b, (T⇄G) = c, (C⇄A) = d, (C⇄G) = e, (A⇄G) = f GTR is parametrized such that f = 1;
F84 [93]	$\pi_T \neq \pi_C \neq \pi_A \neq \pi_G$ (T⇄C)=$1 + K/(\pi_T + \pi_C)$ (A⇄G)=$1 + K/(\pi_A + \pi_G)$ b = c = d = e = 1;
HKY85 [40]	$\pi_T \neq \pi_C \neq \pi_A \neq \pi_G$ (T⇄C) = K (A⇄G) = K b = c = d = e = 1;
TN93 [92]	$\pi_T \neq \pi_C \neq \pi_A \neq \pi_G$ (T⇄C) = K_1 (A⇄G) = K_2 b = c = d = e = 1;
JC69 [17]	$\pi_T = \pi_C = \pi_A = \pi_G, a = b = c = d = e = f = 1$

Notes:

- All the above models include the continuous variation of substitution rate among sites, which corresponds to the more familiar notation: GTR+Γ, TN93+Γ, HKY85+Γ, F84+Γ, JC69+Γ. There is no option of a proportion of "invariable" sites in BYPASSR.

- Comprehensive reviews of the DNA substitution model can also be found in *Computational Molecular Evolution* [48], *Inferring phylogenies* [58], *Molecular systematics* Ch.11, [19] or an review article *Model selection in Phylogenetics* [134].

- **<burnin>** This option allows you to run the burn-in for a given number of iterations or for a specified period of time.

  100000 BYPASSR will run burn-in for 100,000 iterations. 100,000 is the miminum number of iterations.

  1.2h BYPASSR will run burn-in for approximately one hour and fifteen minutes.

- **<sampling>** A number of iterations or an approximate period of time are the options for the sampling stage.

  2e+5 BYPASSR will run for 200,000 iterations. 100,000 is the minimum number of iterations.

  1.5h BYPASSR will run for approximately one hour and thirty minutes.

- **<samples>** The number of samples collected after the chain converged is given here. The printing frequency is (automatically) adjusted accordingly. The recommended numbers are from **1000** to **10000**, usually a rounded number. DEFAULT=2000

- **<random initials>** This option tells the program how to initialize the chain.

  YES BYPASSR starts with random values for all the parameters in the chain. A folder called **output** is automatically generated. If folder with this name existed before, it is replaced by a new empty one. DEFAULT.

  NO BYPASSR restarts with the saved values at the end of the previous burn-in run, using the automatically generated file **temp_variables** in the **output** folder. You can try multiple burn-in runs. Set the burnin iterations to 0 if you want start sampling immediately.

- **<start sampling>** You can choose to run the burn-in and stop at the end of it or to do sampling after the burn-in.

  NO This option can be used if you want to try multiple burn-ins and verify convergence. It will stop after burn-in.

  YES Samples from chain are being collected. The printing frequency is automatically calculated such that you gather the number of samples specified at **<samples>**. DEFAULT.

- **<save site rates>** This option refers to the sample points of site-specific rates.

  NO BYPASSR generates automatically summary statistics for site-specific rates without saving the rates. DEFAULT.

  YES It stores the samples of site specific rates for more statistical analysis than provided. The number of samples specified previously are collected at each site generating a large file (in MB).

- **<save ancestral states>** YES to output the posterior probabilities for the nucleotides at the internal nodes. NO is DEFAULT.

- **<output treefile>** BYPASSR outputs the estimated branch lengths in a tree of newick format. Give file name.

- **<simulation>**

  NO The true values of the branch lengths and substitution rates are not known. DEFAULT.

  YES It uses simulated data and requests a file **generated.rates** (two columns with the site and the rate). The initial branch lengths are considered true. Correlation coefficients for branch lengths and site specific rates vs. their Bayesian estimates are calculated.

- **<clock>**

  YES/NO Global molecular clock is (not) assumed. NO is DEFAULT.

- **<coding>**

  NO Input data is non coding DNA. The site rates summary files contain the sites with the highest and the lowest rates. DEFAULT.

  YES BYPASSR outputs site specific rates at the 1st codon position, 2nd codon position and the 3rd one, assuming that the first nucleotide in the aligned DNA sequence is corresponding to the first codon position.

- **<acceptance rate>**

  DEFAULT The tuning parameters for continuous variables in the chain, as described in the **Theory** section, are adjusted such that to maintain an acceptance rate between 0.2-0.6.

  YES The acceptance rate interval can be modified (e.g. 0.15-0.55).

- **<time on branches>**

  DEFAULT Expected number of substitutions per unit time.

SYN	With <**coding**>:YES; Expected number of synonymous substitutions per unit time. The substitution rates at the 3rd codon position are fixed to 1.

Notes:

- The words between < ... > are keywords and cannot be modified by the user.
- What is written between > and # is processed as variables used in the program. The empty spaces can vary. It does not matter if the variable is written with upper or lower case.
- After the # mark you can write your comments. You can also insert new lines with your own comments.
- If you prefer a different order of the lines in the control file you can swap them or even take some of the options you don't want to use. The DEFAULTS are used instead.

Examples control files:

EXAMPLE 1:

<input datafile>	/home/user/data/DNAfile.fasta
<input treefile>	/home/user/data/TREEfile.tre
<seed>	random
<DNA substitution model>	gtr
<burnin>	1.5h
<sampling>	2e+5
<samples>	2000
<random initials>	yes
<start sampling>	yes
<save site rates>	NO
<output treefile>	./output/estimated.tre
<acceptance rate>	DEFAULT

BYPASSR starts with random values for the parameters. It will run one hour and half in burn-in and continues with 200,000 iterations in sampling, without checking for convergence. During sampling, it collects 2000 points for calculating the statistics of the parameters of interest. If you are lucky enough, the convergence was achieved and the results are good representation of the target distributions. Maybe the following hints will also help: a dataset of 20 taxa and 2000 sites requires roughly 800,000 iterations

in burn-in and 800,000 in sampling; the same magic number of iterations should suffice for a dataset of 50 taxa and 750 sites. Also consider that the approximate number of iterations required to update at least once all branch lengths and site-rates is given by the number of sites multiplied by twice the number of sequences.

EXAMPLE 2:

<input datafile>	/home/user/data/DNAfile.fasta
<input treefile>	/home/user/data/TREEfile.tre
<seed>	random
<DNA substitution model>	gtr
<burnin>	100000
<sampling>	1.2h
<samples>	2000
<random initials>	yes
<start sampling>	no
<save site rates>	NO
<output treefile>	./output/estimated.tre
<acceptance rate>	DEFAULT

BYPASSR starts with random initial values for the parameters. It will run for 100,000 iterations in burn-in and then it will stop. At this point, you can check the convergence as explained in the 2.3.8. Then, the program may be restarted from the point where was left (<random initials> NO) for another burn-in run or the sampling starts by setting <burnin> 0 and <start sampling> yes.

A.4.2 Control file BYPASSR-Gen

Change directory to **GenerateDATA**. The two main source files are written with capital letters. One is to generate tree and the other one is to generate tree and sequences.

Generate random tree only

By compiling with

```
g++ CREATE_RANDOM_TREE.cc -o generatetree
```

the executable **generatetree** is generated. Once you initiate the execution, you wil be asked to choose:

- clock= 1 for yes or 0 for no

- - if yes

  -> enter number of taxa=

  -> desired total treelength=

  A file *"...taxa.tre"* with the tree in newick format is created.

- - if no

  -> enter number of taxa=

  -> enter lambda= ;($1/\lambda$ is the average branch length in the tree)

  A file *"...taxa.tre"* with the tree in newick format is created.

Generate random tree with data sequences

First edit the file **generate_control.ctl** and set the parameters:
clock= (yes or no)
- if yes: enter tree length
- if no:
enter lambda=
nrtaxa=
nrsites=
alpha=
pi_T=
pi_C=
pi_A=
pi_G=
a=
b=
c=
d=
e=
topology= (1 or 0)
rates= (1 or 0)

Topology 1 means new topology, while 0 is a tree you have to generate sequences for different number of sites for example. **Rates** set to 1 draws values from a (continuous) Gamma distribution according to the α you set and one rate for each site. The same rates can be used to generate data for different trees by setting **rates** to 0 and copying the file with your rates in a file called **temp_rates** with two columns site and rate at the site. If you do not have such a file run first with **rates** set to 1 and a file **temp_rates** is automatically generated. I have let just GTR DNA substitution model, but if one is interested in other models I can modify the code for different models too.

A.5 Quick Start

- generate **bypassr** using **make**
- copy **bypassr** and **controlfile** to the data folder (dna in fasta and rooted tree in newick).
- modify the **controlfile** as in example 1
- run **bypassr** using **./bypassr**
- check the pdf files, gnuplot generated plots, in the **output** folder and the output_summary file for the run summary. The file name in the **output** folder is specific to the parameter the file contains.

A.6 Output Files

An **"output"** folder is created every time you run bypassr with random initial values. If a folder output exists, it will be deleted automatically. This folder will store files from burn-in stage and sampling stage. The printing frequency in sampling is adjusted such that the number of <samples> written in the **"controlfile"** is collected. In the burn-in period the number of samples is fixed at approximately 1000 and the printing frequency is posted. All the burn-in samples are discarded, so this phase is for guidance in assessing convergence.

A.6.1 Burn-in

During each burn-in run, another folder is created. The **"controlfile"** of that run, a file called **"temp_variables_chain1"** and the **"burnin_chain1"** are saved in this folder called **"run_1_burnin"**.

If the parallel version was used, a file is generated for each chain, creating file ...._chain2 etc. The file **"temp_variables_chain1"** contains the last values of parameters in the current run. In the eventuality you run three times burn-in and you want to restart with the results of the first run, the file ""**temp_variables_chain1"** from **"run_1_burnin"** has to be copied in the upper folder, **"output"**.

However, the important file of the burn-in stage is:

---

burnin_chain1

This file contains approximately 1000 samples of the α (2nd col.), convergence parameter (3rd col.), tree length (4th col.), λ and log likelihood of the tree (5th col.).

```
#   alpha      R         TL
1   1.0830   0.0734   1.7466   10.76   -98664.1874
2   0.9647   0.2378   1.5413   11.43   -78637.4948
3   0.7381   0.8900   1.3224   12.87   -64486.4832
```

The convergence parameter R is an empirical approach to check convergence by comparing the degree of overlapping of site rate distributions at two sites that share the same pattern. Knowing that the sites with the same pattern should have the same distribution, the value of R_{pair} should be close to 1 if convergence is achieved and the two distributions fully overlap. If $R_{pair} = 0$, the distributions have not a single bin in common.

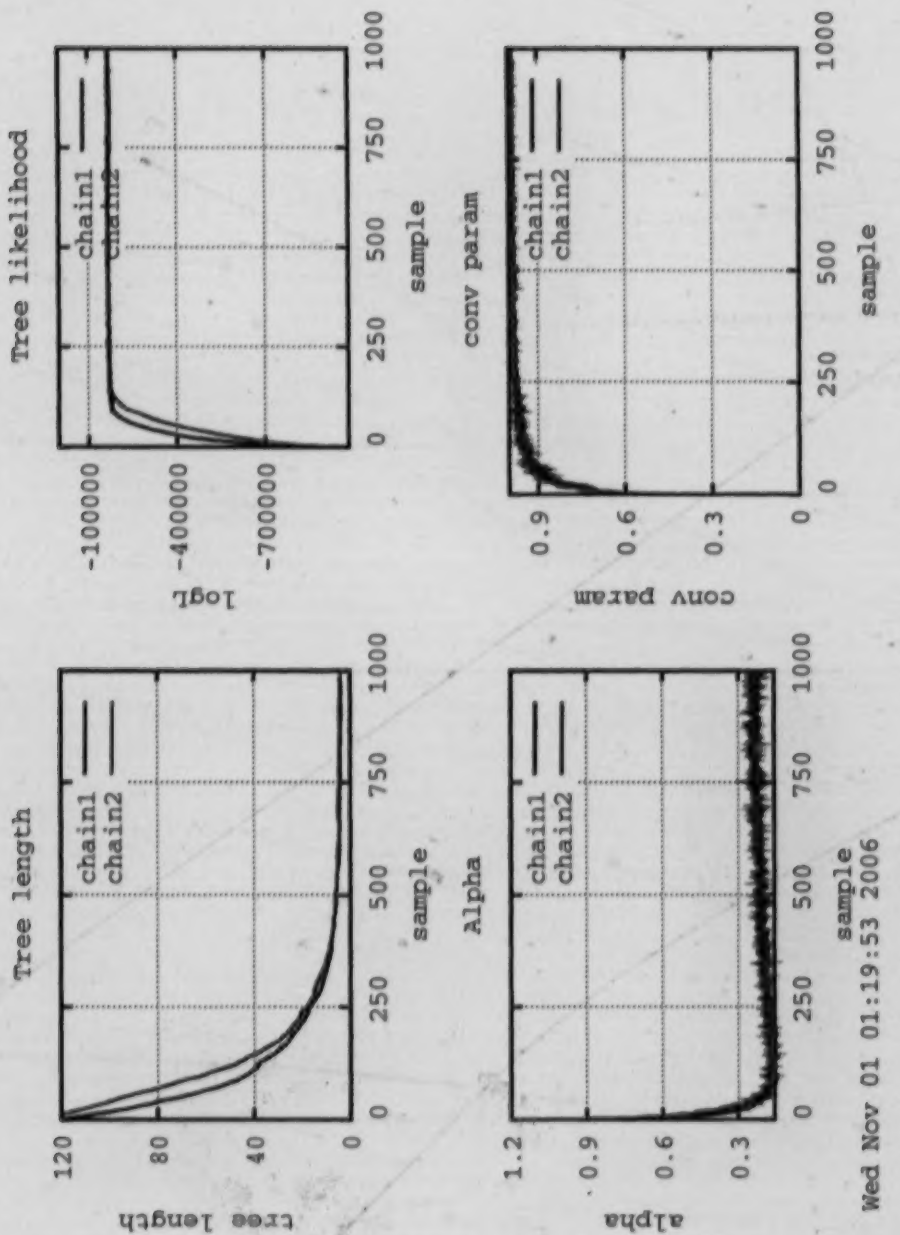
If you have GNUPLOT, the file **"run1_burnin.pdf"** has one page with the four panels as graphical representation of the ***burnin_chain1*** file. The gnuplot script that generated this file is also saved (**"plot-burnin.gnuplot"**). If the pdf file is not created and you know gnuplot is installed, you may run the gnuplot script in terminal "gnuplot plot-burnin.gnuplot". It will use the the "burnin_chain" files from the the folder where the script is.

A.6.2 Sampling

In sampling stage, the program collects samples for the variables of interest. There are several files that are created and located in the **output** folder. Maybe it looks confusing with so many files, but I considered somehow useful to let you choose the summary you are the most interested in and also to give you the possibility to analyze separately the results.

SELECTED_SITES

```
#site nr(align.)        codon        1st       2nd        3rd
   (865-866-867)    codon  289*    1.3940    2.5265     0.3528
   (874-875-876)    codon  292*    1.2144    1.5976     0.2332
   (994-995-996)    codon  332*    2.0238+   2.2662+    0.7094
```


Tree likelihood
chain1
chain2
logL
-100000
-400000
-700000
0
250
500
750
1000
sample
conv param
chain1
chain2
conv param
0.9
0.6
0.3
0
250
500
750
1000
sample
Tree length
chain1
chain2
tree length
120
80
40
0
0
250
500
750
1000
sample
Alpha
chain1
chain2
alpha
1.2
0.9
0.6
0.3
0
250
500
750
1000
sample
Wed Nov 01 01:19:53 2006

```
(1000-1001-1002) codon  334   1.5193    0.0183    0.6160
```

The estimates for the substitution rates are unitless so they make sense in relation with the other rates only. An empirical threshold, H, can be found such that 5% of the analyzed sites at the first position in the codons has the 0.95-quantile above H. These sites are marked as having the rate high and are flagged with '+'. The same is done for the sites located at the second position. At the codon level, if substitution rate at the second position is greater than the one at the first position and has the its posterior mean greater than >1, that codon is listed. A codon is marked with '*' if the substitution rate at its second position exceeds the other two rates within codon.

The candidate sites for negative selection should have the substitution rates close to 0. The 5% of sites located at the first position in the codons, that have the lowest means are marked with '-'. The same is done with 5% of the sites at the second position in codons. Then, the codons with the lowest substitution rates at its first and second codon position marked with '-' are tabulated in the "SELECTED_SITES" file.

If the input sequences are noncoding, the output file contains the 5% of the highest and lowest rates with H calculated in the same way, but across all the sites.

If the option for <time on branches> is YES, the branch lengths are represented in expected number of synonymous substitutions (considering all the substitutions at the third codon position as synonymous). We fix the rate at the third position to 1. Knowing that all the substitutions at the second codon position and 96% of the ones at the first codon position are nonsynonymous, a substitution rate at these sites greater than 1 has the same interpretation as dN/dS rate ratio ω. A site i with substitution rate $r_i > 1$ is a good candidate for being a positively selected site. In the file "SELECTED_SITES" all these sites are printed with their posterior mean rate.

SUMMARY_SITERATES

The number you input in the **"controlfile"**, <samples> option, gives the number of collected samples on which the mean of the posterior distribution is calculated. The first column corresponds to the site number in the original aligned data. The number in the second column is the site number after the data was cleaned. The pattern and how many sites are in that pattern are shown in the following next columns. The sequence at that site is next. The last columns shows the posterior mean substitution rate, its standard deviation (square root of variance) and the mode of the rate at that site. If multiple chains were run, the mean and the standard deviation are averaged over the chains. The estimates for the sites that share the same pattern are summarized by averaging over all of sites in the pattern.

#site #align	#site clean	pat.	data	mean	std.dev	mode
259	1	5	3 GGAGGGGGGG	0.7129	0.5984	0.2500
260	2	1	100 CCCCCCCCCC	0.2295	0.3314	0.0500
261	3	6	1 CCCCTCCCAC	1.3909	1.0480	0.5500
262	4	2	50 TTTTTTTTTT	0.1508	0.2109	0.0500

Note:

The site number in the second column is identical with the site number from the **"rates"** file of **PAML-BASEML** [94].

SUMMARY_SITERATES_HPD

This file contains the same summary statistics as the above file, but the nucleotide sequences are missing. The HPD 95% refers to the highest posterior density, the interval coverage for 95% of the density mass. The narrower the interval, the better meaning smaller variance in the estimate. Multiple intervals shows a very spread distribution or not enough sampled points.

#site #align	site clean	mean	std.dev	mode	HPD\%	
283	1	1.3632	0.1984	0.6500	(0 - 3.7)	
289	2	2.2298	0.3314	1.5500	(0.3 - 4.8)	
301	3	2.6483	0.5480	1.8500	(0.2 - 5.7)	(5.8 - 6)
343	4	2.1525	0.4109	1.5500	(0.3 - 4.4)	(4.5 - 4.6)

Note:

The site number in the second column is identical with the site number from the **"rates"** file of **PAML-BASEML** [94].

SiteSpecificRatesPostDensity

In the source file *global.cc*, **totbinsRATE=200** gives the number of total bins in which the posterior means are distributed based on a **binRATE=0.1** interval increment. The rates above 20 will be all lumped in the 200th bin. A very few data sets had sites with rates greater than 20, requiring more bins. You can modify the source code and increase the number of total bins or just increase the interval to 0.2. Recompile the source code with **make** and reanalyze your dataset.

The output file stores the posterior probability density of substitution rate at each site and it is impossible to be read unless you pick just the column corresponding to the site you are interested in. GNUPLOT is a nice option to plot the distribution at that site. The problem this time is that I cannot create a script to include the posterior probabilities for each site. If you are interested about a particular site (or more sites), write the the following text in a **script-gnuplot.txt** in the **output** folder.

```
set te post land enh color
set out "file.ps"
set xlabel "rate, binned"
set ylabel "frequency"
set title "Post Prob distribution of substitution rate at site 10"
plot "Site-specificRatesPostDensity" u 1:10 t "Post prob at site 10" w boxes
fill solid
set title "Post Prob distribution of substitution rate at site 50"
plot "Site-specificRatesPostDensity" u 1:50 t "Post prob at site 50" w boxes
fill solid
set te X11
!ps2pdf file.ps
!rm file.ps
```

Then, type in terminal

```
gnuplot script-gnuplot.txt
```

This way a file with two pages (one page for each site) is created **file.pdf** and the posterior probabilities are shown with bars of width 0.1. The values on x axis correspond to the substitution rate and on y is the frequency of that rate among the collected sample. The site number here is the one as in cleaned data (see SUMMARY_SITERATES file for the corresponding number in the alignment with gaps).

siterates_chain1

This file is saved only if you write to <save site rates> YES. The values in it were the source of all the statistics in the previous files description.

SUMMARY_BRLENS

This is the only file to summarize branch lengths estimates. The internal branches, the branches corresponding to the tips are specified accordingly. The second column shows the branches at the root. The last two columns are the posterior mean and the standard deviation for each of the branch length estimate.

```
int_branch          -          1        0.4843    0.0044
int_branch       root_br       2        0.0636    0.0004
SEQUENCE_NAME       -          3        0.2545    0.0023
```

---

usertreebrlens

This file stores the branch lengths from the provided tree in the node order used in BYPASSR. If you want to graphically compare them with the estimated lengths from BYPASSR, use the following script.

```
set te post land enh color
set out "brlens-comparison.ps"
set title "Estimated vs. initial branch lengths"
plot "SUMMARY_BRLENS" u 3:4 t "bypassr" w lp lw 2, "usertreebrlens" u
1:2 t "initial"w lp lw 2
set te X11
!ps2pdf brlens-comparison.ps
!rm brlens-comparison.ps
```

Then, type in terminal

```
gnuplot script-gnuplot.txt
```

---

treelength_chain1, logL_chain1, alpha_chain1, lambda_chain1

The samples of treelength, alpha, lambda or tree loglikelihood along the chain are saved. All these four files are summarized as plots in the file **"sampling.pdf"**. The gnuplot script that generated this file is also saved (**"plot-sampling.gnuplot"**).

---

ancestral_chain1, ANCESTRAL_STATES